Quantitative topics in portfolio and risk management

by Emlyn Flint

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor

University of Pretoria

Faculty of Natural and Agricultural Sciences
Department of Actuarial Science, and Department of Mathematics and Applied Mathematics

Supervisor: Professor Eben Maré

14 July 2019

Emlyn Flint
Quantitative topics in portfolio and risk management
Quantitative Finance, 14 July 2019
Supervisor: Professor Eben Maré

University of Pretoria
Faculty of Natural and Agricultural Sciences
Department of Actuarial Science, Department of Mathematics and Applied Mathematics
Mathematics building, University of Pretoria, Corner of Lynnwood Road and Roper Street
Pretoria, 0028

To Laura

Abstract

The modern quantitative portfolio manager is the quintessential "jack of all trades". Not only do they need to be an expert in the specific area of portfolio management, they also need to have a thorough understanding of the related areas of valuation, data processing, risk management and performance analysis. What this means practically is that quantitative portfolio managers are regularly faced with problems spanning the entire $\mathbb{P} - \mathbb{Q}$ spectrum of quantitative finance. Spurred by this reality, the central research question motivating this thesis is exactly the core motivation behind every decision taken by a quantitative portfolio manager: *What is the most efficient, practical method for constructing, managing and evaluating optimal multi-asset portfolios in dynamic, non-normal markets?* In this thesis, we attempt to provide insight into this broad central research question by offering new perspectives and practical solutions to a selection of sub-problems that a quantitative portfolio manager would have to address in practice. In particular, this thesis is comprised of six essays that each tackle specific problems in the related areas of derivatives, return modelling, systematic trading strategies and portfolio construction.

Acknowledgements

This thesis would not have happened without a great supporting cast.

Eben: for invaluable guidance, patience, dedication and motivation. I couldn't have asked for a better supervisor. You always knew just the right amount of carrot and stick needed to get me to the finishing line.

Laura: for putting up with me over the last four years and allowing me the freedom to pursue this passion.

Anthony, Chris, Edru, Florence, Kanshu, Kovlin and Simon: colleagues, co-authors and friends. You were invaluable in your contributions and feedback throughout all the research.

Examiners, reviewers, conference attendees and editorial staff: for all the feedback and suggestions I received during the various submission processes and presentations. Your thoughts, ideas and suggestions were invaluable in improving this research.

Family and friends: for all the love, support, stress, frustration, some more support, and sound advice given.

Legae Peresec: for providing several invaluable datasets.

LaTeX Stack Exchange community: if not for you, this thesis would never have made it past page 1. You saved me weeks of typesetting frustration.

As always, all errors in this manuscript are my own.

Contents

List of Figures

List of Tables

List of Acronyms

ALSI	FTSE/JSE All Share Index
APT	arbitrage pricing theory
ATM	at-the-money (option / volatility)
BP	beta parity (portfolio)
BSM	Black-Scholes-Merton (model)
BtM	book value to market value
CAPE	cyclically adjusted price-to-earnings (ratio)
CAPM	capital asset pricing model
CLI	composite leading index
CW	capitalisation weight (portfolio)
DD	drawdown
ERC	equal risk contribution (portfolio)
EW	equal weight (portfolio)
fBm	fraction Brownian motion (process)
FBSI	fractional Black-Scholes inspired (model)
FER	factor efficiency ratio
FLOAM	fundamental law of active management
FTSE	Financial Times Stock Exchange
GARCH	generalised autoregressive conditional heteroscedasticity
GDP	gross domestic product

GMV	global minimum variance (portfolio)
HMM	hidden Markov model
JSE	Johannesburg Stock Exchange
LFM	linear factor model
MA	moving average
MDP	maximum diversification portfolio
MPT	modern portfolio theory
MV	minimum variance (portfolio)
NCD	negotiable certificate of deposit
OLS	ordinary least squares
OTC	over the counter (derivative)
OTM	out-the-money (option / volatility)
PDE	partial differential equation
RBSA	returns-based style analysis
RND	risk-neutral distribution
RP	risk parity (portfolio)
SA	South Africa
SABR	stochastic alpha beta rho (model)
SARB	South African Reserve Bank
SVI	stochastic volatility inspired (model)
Swix40	FTSE/JSE Shareholder Weighted Top40 Index
TAA	tactical asset allocation
TE	tracking error
Top40	FTSE/JSE Top40 Index
TPM	transition probability matrix
USDZAR	United States Dollar to South African Rand (exchange rate)
VaR	value-at-risk

Part I

The State of Modern Quantitive Finance

1 Introduction

> *It's not easy taking my problems one at a time when they refuse to get in line.*
>
> — **Ashleigh Brilliant**
> (Author and cartoonist)

1.1 Modern Quantitative Finance

What is quantitative finance? It may surprise one to find that there is no universally accepted answer to this question. Emmanuel Derman, one of the pioneering quants on Wall Street, wrote the following on the subject in his 2004 biographical account, *My Life as a Quant*:

> Quants and their cohorts practice "financial engineering" — an awkward neologism coined to describe the jumble of activities that would better be termed *quantitative finance*. The subject is an interdisciplinary mix of physics-inspired models, mathematical techniques, and computer science, all aimed at the valuation of financial securities.

We can make two observations from this definition. Firstly, quantitative finance is multi-disciplinary by its very nature. Derman already mentions the fields of mathematics, computer science and physics. Additional to this are the ideas stemming from finance, statistics, economics and, more recently, data science. Because of this broad mixture, quantitative finance is also commonly referred to as mathematical finance, computational finance, financial engineering, mathematics of finance, or quantitative economics. The second observation is that Derman suggests that *valuation* is the primary subject of quantitative finance. This surely reflects the time of his entry into the financial industry in the 1980s; a time when formal option pricing theory was still in its infancy, physicists were moving en masse from academia into finance, and new derivative contracts were constantly being designed and traded. Paul Wilmott, another leading quant and contemporary of Derman's, gives a similar definition in his seminal textbooks on derivatives valuation and risk management, referring to these subjects as *classical quantitative finance*.

As markets have progressed though and computational power has become increasingly cheap, the use of quantitative methods and techniques has spread to all areas of finance. As a result, quantitative finance today encompasses far more than just derivatives valuation. In order to understand the complete spectrum of *modern quantitative finance*, we consider below the comprehensive, modular framework developed by Meucci (2019).

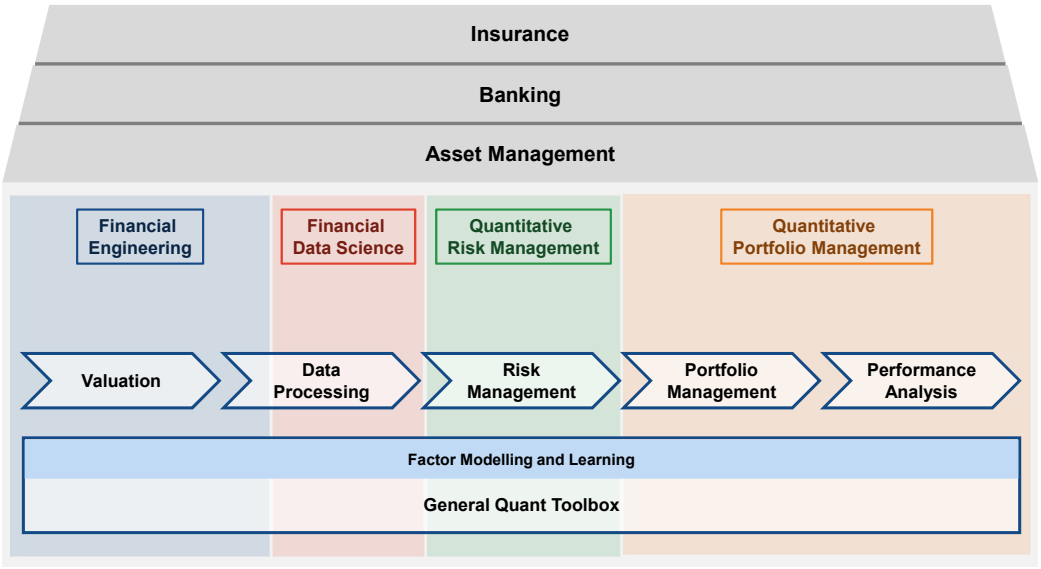


Figure 1.1. An overview of modern quantitative finance with respect to core business sectors, financial functions and specialities. Figure adapted from Meucci (2019).

Figure 1.1 displays the core business sectors, financial functions and specialities within modern quantitative finance. The financial industry can largely be split into three sectors: asset management, banking and insurance. Common to all of these sectors are the five broadly sequential financial functions: current valuation, data processing, ex ante risk management and portfolio management, and ex post performance analysis.

One can also break these functions into four separate quantitative specialisations. *Financial engineering* remains the term most associated with derivatives valuation and risk management. Given the focus on valuation, it also includes an aspect of data processing, usually in the form of stochastic modelling. *Financial data science* is perhaps the newest specialisation, focussing for now on general data processing. This may well grow in future though to encompass other financial functions. *Quantitative risk management* focusses on modelling and estimating the probability of future scenarios occurring, and thereafter attributing the consequences thereof across the relevant assets. *Quantitative portfolio management* covers both the financial and mathematical aspects of optimal portfolio construction, trade execution and performance attribution.

Lastly, there exists the set of general quantitative techniques used across all sectors, functions and specialities. Within this toolbox, special mention should be made of (non)linear factor models and their associated machine learning techniques, which are predominant in almost all quantitative finance applications.

1.1.1 The $\mathbb{P} - \mathbb{Q}$ Spectrum of Modern Quantitative Finance

Another way of understanding quantitative finance is to categorise it by the type of probability measure used. In particular, risk and portfolio management are concerned with estimating, modelling and ultimately managing the probability of future real-world events. This set of probabilities is denoted by $\mathbb{P}$. In contrast, the derivatives valuation models based on the theory of arbitrage make use of a different probability measure, referred to as risk-neutral and denoted by $\mathbb{Q}$. Table 1.1 summarises the major differences between the real-world $\mathbb{P}$ and risk-neutral $\mathbb{Q}$ sections of quantitative finance.

Table 1.1. Differences between $\mathbb{P}$ and $\mathbb{Q}$ quantitative finance. Adapted from Meucci (2019).

	Risk & Portfolio Management	**Derivatives Pricing**
Goal	Forecast the future	Extrapolate the present
Environment	Real-world probability $\mathbb{P}$	Risk-neutral probability $\mathbb{Q}$
Processes	Discrete-time series	Continuous-time martingales
Dimension	Large	Small
Tools	Multivariate statistics	Itô calculus, PDEs
Challenges	Estimation	Calibration
Business	Buy-side	Sell-side

$\mathbb{P}$ quants focus on risk and portfolio management. In essence, they use data and quantitative models to estimate future asset return distributions, and then use this information within an investment process to make decisions that will optimise the investor's future profit-and-loss profile. Data processing and estimation are thus key aspects of $\mathbb{P}$ modelling, especially in light of the large dimensionality faced by most portfolio managers.

$\mathbb{Q}$ quants focus on derivatives pricing and risk management. As a result, $\mathbb{Q}$ modelling makes use of stochastic calculus and partial differential equations (PDEs) to calculate the present value of a derivative contract under current conditions and across various market scenarios, and in the presence of counterparty credit risk. As we mentioned above, quants were historically focussed on the niche area of $\mathbb{Q}$ modelling, driven largely by market demand at the time as well as the skill set of the incoming physicist-turned-quants.

In recent years though, there has a been a resurgence of interest in the considerably broader area of $\mathbb{P}$ quantitative finance and a concurrent decrease of interest in the pure risk-neutral $\mathbb{Q}$ space.[1] This has led to the rise of a new type of $\mathbb{P}$ quant. It is in this milieu that we present our current body of research.

1.2 Thesis Scope and Aims

The modern quantitative portfolio manager is perhaps the quintessential "jack of all trades". This is not meant in a derogatory manner. Not only does the manager need to be an expert in the area of *quantitative portfolio management*, they also need to have a thorough understanding of all of the financial functions outlined in Figure 1.1. This includes even the *valuation* function because, ultimately, the manager will always need to decide whether the current price of an asset is fair given their view on its future return distribution. What this means practically is that modern quantitative portfolio managers are faced with problems spanning the entire $\mathbb{P} - \mathbb{Q}$ spectrum of quantitative finance.

Motivated by this reality, the objective of this thesis is to provide new perspectives and practical solutions to a selection of problems across the $\mathbb{P} - \mathbb{Q}$ spectrum that a quantitative portfolio manager would perforce have to address. To this end, Table 1.2 provides a summary of the journal articles, conference proceedings and working papers stemming from the entirety of this research agenda. This thesis then comprises a selection of six essays from this body of work, each presented in separate, independent chapters. While each chapter can be read and assimilated in isolation, when taken in conjunction, they form a larger holistic contribution aimed at better understanding multiple theoretical and empirical aspects over all of the five modern quantitative finance functions.

1.2.1 Chapter Overview

The essays presented in this thesis are categorised into two parts. Part II is made up of Chapters 2, 3 and 4 and deals with topics predominantly from the $\mathbb{Q}$ side of the quantitative finance spectrum. Part III includes Chapters 5, 6 and 7 and covers topics from the $\mathbb{P}$ side of the spectrum.

In Chapter 2, we address several theoretical and practical issues in option pricing and implied volatility calibration in a fractional Black-Scholes market. In particular, we discuss how the fractional Black-Scholes model admits a non-constant implied

[1] The new field of X-Value Adjustment (XVA) is a pertinent example of how risk-neutral derivatives valuation is adjusted to account for real-world issues such as credit risk and capital requirements.

Table 1.2. Published doctoral research: journal articles, conference proceedings and online working papers.

Peer-reviewed publications

1. **Flint, E.**, & du Plooy, S. (2018). Extending Risk Budgeting for Market Regimes and Quantile Factor Models. *Journal of Investment Strategies*, 7(4), 51-74.
2. Perumal, K., & **Flint, E.** (2018). Systematic Testing of Systematic Trading Strategies. *Journal of Investment Strategies*, 7(3), 29-49.
3. **Flint, E.**, & Maré, E. (2017). Estimating option-implied distributions in illiquid markets and implementing the Ross recovery theorem. *South African Actuarial Journal*, 17(1), 1-28.
4. **Flint, E.**, Seymour, A., & Chikurunhe, F. (2017). Factor Investing in South Africa. *Alternative Investment Analysts Review*, 6(2), 19-36.
5. **Flint, E.**, & Maré, E. (2017). Fractional Black-Scholes option pricing, volatility calibration and implied Hurst exponents in South African context. *South African Journal of Economic and Management Sciences*, 20(1), 1-11.
6. Baker, C., Rajaratnam, K., & **Flint, E.**. (2016). Beta estimates of shares on the JSE Top 40 in the context of reference-day risk. *Environment Systems and Decisions*, 36(2), 126-141.

Conference proceedings (excluding published articles)

1. **Flint, E.**, Seymour, A., & Chikurunhe, F. (2017). Regime-based tactical allocation for equity factors and balanced portfolios. *Actuarial Society 2017 Convention*.
2. Wessels, J., & **Flint, E.** (2017). Alternative and new methods for measuring mutual fund performance. *Southern African Finance Association 2017 Conference*.

Working papers (excluding published articles and coference proceedings)

1. **Flint, E.**, Seymour, A., & Chikurunhe, F. (2018). Estimation with Flexible Probabilities: Measuring Rand Hedges, Finding Diversifiers, Enhancing Style Analysis, *SSRN Electronic Journal 3324076*.
2. Seymour, A., **Flint, E.**, & Chikurunhe, F. (2018). Dynamic Portfolio Management Strategies: A Framework for Historical Analysis, *SSRN Electronic Journal 3214453*.
3. **Flint, E.**, Seymour, A., & Chikurunhe, F. (2017). Defining Activeness: Active Share, Risk Share & Factor Share, *SSRN Electronic Journal 2767436*.
4. **Flint, E.**, Chikurunhe, F., & Seymour, A. (2015). The Cost of a Free Lunch: Dabbling in Diversification, *SSRN Electronic Journal 3084637*.
5. **Flint, E.**, Seymour, A., & Chikurunhe, F. (2015). In Search of the Perfect Hedge Underlying, *SSRN Electronic Journal 2767532*.

volatility term structure when the Hurst exponent is not 0.5, and that one-year implied volatility is independent of Hurst exponent and equivalent to fractional volatility. Building on these observations, we introduce a novel eight-parameter fractional Black-Scholes inspired, or FBSI, model. This deterministic volatility surface model is based on the fractional Black-Scholes framework and uses Gatheral's (2004) stochastic volatility inspired (SVI) parameterisation for the fractional volatility skew and a quadratic parameterisation for the Hurst exponent skew. We address the issue of arbitrage-free calibration for the FBSI model in depth and prove in general that any FBSI volatility surface is free from calendar-spread arbitrage. We then test the FBSI model empirically on implied volatility data on a South African equity index as well as the USDZAR exchange rate. Results show that the FBSI model fits the equity index implied volatility data very well and that a more flexible Hurst exponent parameterisation is needed to accurately fit the USDZAR implied volatility surface data.

In Chapter 3, we describe how forward-looking information on the statistical properties of an asset can be extracted directly from options market data and how this can be used practically in portfolio management. Although the extraction of a forward-looking risk-neutral distribution is well-established in the literature, the issue of estimation in an illiquid market is not. We use the deterministic SVI volatility model to estimate weekly risk-neutral distribution surfaces. We consider the issue of calibration with sparse and noisy data at length and propose a simple but robust fitting algorithm. Furthermore, we extract real-world implied information solely from options data by implementing the recovery theorem introduced by Ross (2015). Recovery is an ill-posed problem that requires careful consideration. To this end, we describe a regularisation methodology for extracting real-world implied distributions and implement this method on a history of equity index SVI implied volatility surfaces. We analyse the first four moments from the implied risk-neutral and real-world implied distributions and use them as signals within a simple tactical asset allocation framework, finding promising results.

In Chapter 4, we attempt to answer the question: *What underlying portfolio should one use to hedge an active fund*? We start by considering three different decompositions for active return and tracking error respectively, in order to have a complete understanding of the underlying sources of risk and reward in the fund. We then describe a general mixed integer programming framework that allows us to select a sub-basket of assets that will most accurately replicate those sources of risk and reward whilst simultaneously complying with real-world market constraints. We then study how the effectiveness of an index hedge decreases with tracking error, where effectiveness is measured in terms of the change in downside risk measures of the hedged portfolio. Motivated by these three elements, we introduce several alternative hedging methods for the active fund manager. In particular, we focus

specifically on the use of long-only and long/short custom basket options as a means of creating an appropriate portfolio hedge. A novel pricing methodology for long/short basket options is also introduced.

In Chapter 5, we provide an introduction to, and critique of, the factor investing paradigm in a South African setting. We initially discuss the general factor construction process at length and construct a comprehensive range of risk factors for the South African equity market according to international factor modelling standards. We focus on the size, value, momentum, profitability, investment, low volatility and low beta risk factors respectively. We critically examine the historical behaviour and robustness of these factors, paying particular attention to the issues of long-only versus long/short factors, the impact of size, the effect of rebalancing frequency and date, and the robustness of performance to alternative factor definitions. We also review how these factors can be used generally in risk management and portfolio management. To this end, we consider factor risk attribution and returns-based style analysis in the risk management space, and multi-factor portfolio construction methods in the portfolio management space.

In Chapter 6, we consider whether regimes can add value to the asset allocation process. Four methods for regime identification – economic cycle variables, fundamental valuation metrics, technical market indicators and statistical regime-switching models – are discussed and tested on two asset universes – long-only South African equity factor returns and representative balanced portfolio asset class returns. We find several promising regime indicators and use these to create two regime-based tactical allocation frameworks. Out-of-sample testing on both the equity factor and balanced asset class data shows very promising results, with both regime-based tactical strategies outperforming their respective static benchmarks on an absolute and risk-adjusted return basis.

In Chapter 7, we combine several disparate avenues in the literature to create a novel, unified risk-based optimisation framework. Specifically, we extend the existing risk budgeting approach of Richard and Roncalli (2015) to allow for changing market regimes, factor dependence and nonlinear and asymmetric market structure. We show that the existing framework can be readily extended to include a factor-dependent return process using standard models available in the literature. Structural changes in market conditions are then incorporated into the framework through the use of a regime-switching turbulence index. Finally, a nonlinear and asymmetric market dependence structure is accounted for by using quantile factor models. Most importantly, this extended framework is only comprised of a series of linear models, and is thus simple to understand and implement. We consider two applications of the extended framework, namely scenario analysis and parameter uncertainty analysis, through means of a simple empirical case study. Finally,

we introduce the concept of Risk Maps, which provide managers with a graphical approach for estimating and evaluating risk optimality in a multi-objective and multi-scenario setting.

1.2.2 Links with the Modern Quantitative Finance Framework

In addition to categorising chapters by where they separately lie on the $\mathbb{P} - \mathbb{Q}$ spectrum, it is also important to understand how each chapter overlaps and complements each other. This is most easily done in the context of the underlying quantitative finance framework. Figure 1.2 categorises the chapters based on their relation to the five functions – or areas if you will – of modern quantitative finance.

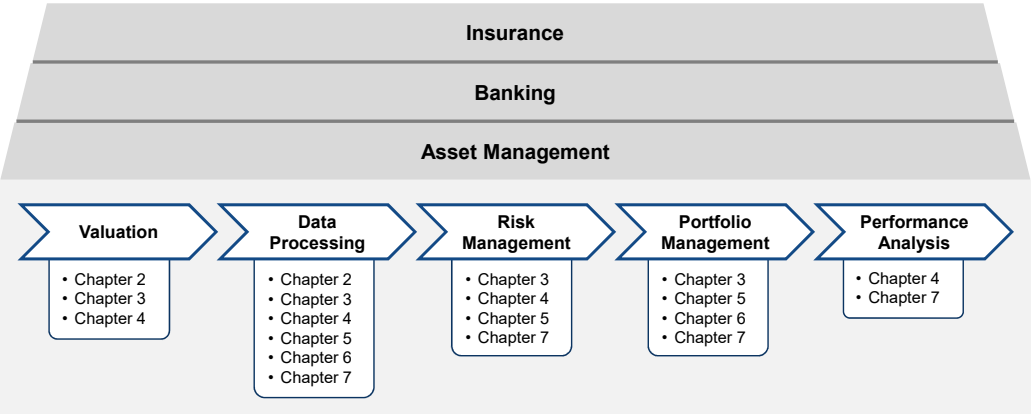


Figure 1.2. Thesis chapters classified across modern quantitative finance functions.

Starting with the valuation function, it is no surprise to see that the three $\mathbb{Q}$-based chapters fall under this area. However, each chapter deals with slightly different but related aspects of derivatives valuation. In Chapter 2, we consider how to model the implied volatility surface – and thus ultimately value contingent claims – in a market which is driven by a non-standard fractional brownian motion process. Chapter 3 starts by addressing a similar modelling issue, applying an existing deterministic implied volatility but in an illiquid market setting. It then diverges by considering how to transform the risk-neutral implied volatility surface into a real-world valuation of the ex ante market risk premium. Chapter 3 is thus a good example of a chapter that has its roots in a $\mathbb{Q}$-based problem but is ultimately applied in a $\mathbb{P}$ setting. Chapter 4 then considers the use of derivatives in an actual portfolio setting and also introduces a novel pricing methodology for long/short basket options.

All chapters are included in the data processing function but deal with different forms of data or different types of processing. Chapter 2 assumes the underlying market follows a fractional brownian motion and uses this assumption to model the implied

volatility surface separately as a fractional volatility component and an implied memory component. Chapter 3 also models the implied volatility surface through time but then uses results from multivariate statistics to transform these surfaces into forward-looking real-world probability distributions. Chapter 4 uses simulation to estimate the real-world effectiveness of various hedging structures under different market conditions. It also introduces a method to simulate increasingly active portfolio weight distributions in a complete manner. Chapter 5 is ultimately a large-scale exercise in linear factor modelling for identifying the true asset return drivers and thus estimating better expected returns. Chapter 6 assumes that markets follow a regime-switching process and considers several methods for identifying, estimating and ultimately forecasting these regimes. Finally, Chapter 7 combines several of the approaches used in previous chapters and applies them in a risk budgeting setting. Specifically, markets are modelled using regime-switching, nonlinear factor models. Although methods may differ across the chapters, the ultimate goal always remains the same: to find the best method for understanding current markets and forecasting their future distributions.

There are four chapters with application in risk management. Chapter 3 estimates future real-world distributions from option markets, which is exactly the first half of what quantitative risk managers focus on. Chapter 4 addresses the effectiveness of several hedging strategies under varying market conditions and thus risk management is the core focus. Chapter 5 discusses how linear factor models are used in risk management and showcases several applications. Chapter 7 builds on this idea through the concept of factor-based risk budgeting. This is essentially a portfolio management philosophy that relies solely on risk management principals in order to construct risk-optimal investment portfolios.

We then have four chapters concerned with the portfolio management function. Chapter 3 considers a switching asset allocation strategy based on real-world distribution forecasts estimated from option markets. Chapter 5 is a direct critique of the factor investing paradigm in an illiquid emerging market setting and thus is directly linked to multiple portfolio management applications. Chapter 6 again considers asset allocation strategies but this time focusing on several methods for implementing tactical views in markets assumed to follow a regime-switching process. Lastly, Chapter 7 focusses on risk budgeting which, as mentioned above, is ultimately an amalgamation of risk management and portfolio management into a single investment process.

Finally, we have two chapters that fall under the performance analysis function, which is by nature an ex post exercise. Chapter 4 analyses the effectiveness of several index hedging strategies under various market conditions for portfolios of increasing activeness. The goal of this analysis is to give managers a scale of effectiveness for

certain index hedge overlays versus the level of portfolio activeness to that index. Lastly, Chapter 7 introduces the idea of Risk Maps, which provide managers with a graphical approach for estimating and evaluating risk optimality in a multi-objective and multi-scenario setting. These can either be used in an ex post performance analysis setting or an ex ante portfolio management setting.

1.2.3 Research Questions

The central research question driving this thesis is:

> What is the most efficient, practical method for constructing, managing and evaluating optimal multi-asset portfolios in dynamic, non-normal markets?

Clearly, this is a very broad question, and definitely one that is too broad to answer in a single thesis. In fact, it is quite likely that this question will never be answered to satisfaction in its entirety. However, that doesn't change the fact that it remains the core motivation behind every decision that a quantitative portfolio manager makes. It is therefore worth framing any quantitative finance research undertaken in terms of its overall impact on this central research question. In this thesis, we attempt to provide insight into this central research question by considering a series of sub-research questions in particular areas of quantitative finance; namely, derivatives, return modelling, systematic trading strategies and portfolio construction.

1. Derivatives research questions:

    a) How do we price and risk-manage derivatives in dynamic, non-normal markets?

    b) How do we practically model the implied volatility surface in illiquid markets with sparse data?

    c) How do we use derivatives effectively as risk management tools in portfolios?

    d) What underlying basket should we use when hedging an active portfolio?

2. Return modelling research questions:

    a) How can we practically extract real-world, forward-looking asset information from derivative markets?

b) Does the factor investing paradigm work in the South African emerging equity market setting?

c) Can regime-switching models be used to forecast future asset return distributions?

d) How do we allow for changing market regimes and nonlinear, asymmetric factor dependencies in the return modelling process in a practical and parsimonious manner?

3. Systematic trading strategies research questions:

a) Can forward-looking option-implied information be successfully used as signals in a systematic trading strategy?

b) Are systematic strategies based on fundamental equity factors profitable and robust in the South African emerging market setting?

c) Can we create regime-based tactical allocation strategies that outperform their respective static benchmarks over the long run?

4. Portfolio construction research questions:

a) How do we effectively construct hedging portfolios to manage downside risk in increasingly active funds?

b) What methods are available for practically constructing optimal single-factor and multi-factor equity portfolios?

c) How do we incorporate dynamic factor dependencies and non-normal markets into the risk budgeting framework?

d) How can we estimate and evaluate portfolio risk optimality in a multi-objective and multi-scenario setting?

The chapters that follow present some new perspectives and practical solutions to this diverse set of research questions.

Part II

Chapters in the $\mathbb{Q}$-Spectrum

2

Fractional Black-Scholes Option Pricing, Volatility Calibration and Implied Hurst Exponents

> *I have abandoned my search for truth and am now looking for a good fantasy.*
>
> — **Ashleigh Brilliant**
> (Author and cartoonist)

Chapter Synopsis

In this chapter, we address several theoretical and practical issues in option pricing and implied volatility calibration in a fractional Black-Scholes market. In particular, we show that the fractional Black-Scholes model admits a non-constant implied volatility term structure when the Hurst exponent is not 0.5, and that one-year implied volatility is independent of Hurst exponent.

Building on these observations, we introduce a novel eight-parameter fractional Black-Scholes inspired, or FBSI, model. This deterministic volatility surface model is based on the fractional Black-Scholes framework and uses Gatheral's (2004) SVI parameterisation for the fractional volatility skew and a quadratic parameterisation for the Hurst exponent skew. The issue of arbitrage-free calibration for the FBSI model is addressed in depth and it is proven in general that any FBSI volatility surface is free from calendar-spread arbitrage.

The FBSI model is empirically tested on implied volatility data on a South African equity index as well as the USDZAR exchange rate. Results show that the FBSI model fits the equity index implied volatility data very well and that a more flexible Hurst exponent parameterisation is needed to accurately fit the USDZAR implied volatility surface data.

This chapter is adapted from the journal article by Flint and Maré (2017b) and addresses research questions 1a and 1b given in Section 1.2.3.

2.1 Introduction

Contingent claims on underlying assets are typically priced under the framework introduced by Black and Scholes (1973). This framework assumes, inter alia, that the log returns of an underlying asset are normally distributed. However, many researchers have shown that this assumption is violated in practice. Cont (2001) put forth several 'stylised facts' of empirical asset returns, defined as 'statistical properties found to be common across a wide range of instruments, markets and time periods'. These include the properties of the so-called heavy tails, volatility clustering, leptokurtosis and long memory. While many authors have studied the implications of these stylised facts across a variety of market applications (see Meucci, 2019, and the references contained therein), this research addresses an issue which has previously not received much attention. More specifically, this research considers several theoretical and practical issues in the pricing of contingent claims when the underlying is assumed to display long memory.

Hurst (1951) proposed a statistical metric – and its estimation methodology – for measuring the long-term memory embedded within a given system. This metric is now commonly referred to as the Hurst exponent, index or parameter and is denoted by $H \in [0, 1]$. For a given time series, $H < \frac{1}{2}$ implies that the series displays a negative, long-term autocorrelation (or anti-persistence), $H = \frac{1}{2}$ implies zero long-term autocorrelation and $H > \frac{1}{2}$ implies that the series displays a positive autocorrelation. In financial calculus parlance, this would be equivalent to a stochastic process respectively displaying mean-reversion, no memory or momentum. A stochastic process with high $H > \frac{1}{2}$ will also be smoother than the same process with low $H < \frac{1}{2}$ because it is less likely to move against the underlying trend.

Mandelbrot and Van Ness (1968) were the first researchers to suggest the use of the Hurst exponent in financial markets. Specifically, they suggested that financial asset prices displayed some form of long memory and introduced fractional Brownian motion (fBm) – a new class of Gaussian random functions – for modelling the log increments in asset price processes. The fBm for a given Hurst exponent H (see, e.g. Biagini et al., 2008) is the continuous Gaussian process $\{B_H(t), t \in \mathbb{R}^+\}$, with

$$\mathbb{E}\left[B_H(t)\right] = 0 \tag{2.1}$$

$$\mathbb{E}\left[B_H(t), B_H(s)\right] = \frac{1}{2}\left[t^{2H} + s^{2H} - |t-s|^{2H}\right]. \tag{2.2}$$

From Equation 2.2, it is clear that the standard Brownian motion is simply a special case of fBm where $H = \frac{1}{2}$. For all other values of H though, the fBm process will have dependent increments. Mandelbrot (2013), as well as the references contained

therein, provides an excellent summary of the early applications of the fBm theory in financial markets. A sample of the more recent studies is given below.

Karuppiah and Los (2005) consider the long-term dependence of Asian currencies, finding empirical Hurst exponents between 0.3 and 0.5 and thus implying antipersistent behaviour. In contrast, they note that equities typically exhibit persistent behaviour, with Hurst exponents estimated between 0.6 and 0.7 (see also, e.g. Peters, 1989, 1994). Simonsen (2003) demonstrates that Nordic electricity spot prices can be modelled using fBm with a Hurst exponent of approximately 0.4. Alvarez-Ramirez et al. (2002) conclude that crude oil price formations are stochastically persistent with long-term memory processes at work. Long-term dependence – as well as heavy-tailed distributions – in high frequency financial data has also been established by Andersen and Bollerslev (1997) and Müller et al. (1998). More recent work by Tzouras et al. (2015) employs the Hurst exponent to model memory-dependent properties in share indices and oil prices (see also Alvarez-Ramirez et al., 2008; Serinaldi, 2010). The Hurst exponent is also used by Cajueiro and Tabak (2004), Jefferis and Thupayagale (2008), Morris et al. (2009), and Rejichi and Aloui (2012) to test the evolving efficiency of emerging equity markets.

Hu and Øksendal (2003) derived closed-form solutions for contingent claim valuation in a fractional Black–Scholes market, where the standard Brownian motion in the asset price process is replaced with an fBm (see also Necula, 2002). Their work was extended by Elliott and Van der Hoek (2003). Specifically, for a market with a risk-free asset A and a risky stock S, a fractional Black–Scholes market is defined as

$$dA(t) = rA(t)\,dt \;\; s.t.\; A(0) = 1;\, r > 0 \tag{2.3}$$

$$dS(t) = \mu S(t)\,dt + \sigma S(t)\,dB_H(t) \;\; s.t.\; S(0) = s > 0;\, \sigma > 0, \tag{2.4}$$

where $0 \leq t \leq T$, r and μ are constant drift parameters and σ is a constant scale parameter. From this, Hu and Øksendal (2003) derive the fractional Black–Scholes value of a European call option $C_f(\cdot)$ at time t with strike K and term $\tau = T - t$ as

$$C_f(S_t, K, \tau, r, \sigma, H) = S_t \Phi\left(\hat{d}_1\right) - Ke^{-r\tau}\Phi\left(\hat{d}_2\right), \tag{2.5}$$

where Φ is the standard cumulative normal distribution function and

$$\hat{d}_1 = \frac{\ln\left(\frac{S_t}{K}\right) + r\tau + \frac{1}{2}\sigma^2\tau^{2H}}{\sigma\tau^H} \tag{2.6}$$

$$\hat{d}_2 = \hat{d}_1 - \sigma\tau^H. \tag{2.7}$$

As with the seminal Black–Scholes option pricing formula (Black & Scholes, 1973), one can infer the valuation formula for a European put option $P_f(\cdot)$ with strike K

and term τ via put-call parity. Furthermore, a dividend yield q can be added to the above equations in a similar manner to Merton's (1973b) extension of the standard Black–Scholes framework.

Although already stated above, Equation 2.5 makes it clear that setting $H = \frac{1}{2}$ simply gives one the classical Black and Scholes (1973) option pricing formula. Therefore, assuming that the risk-free rate and dividend yield are known, fBm option prices are fully described by two parameters: the Hurst exponent H as a measure of long memory and the volatility of the stock σ after controlling for long memory.

The rest of this chapter is organised as follows. Section 2.2 is devoted to the links between standard Black-Scholes volatility and fractional Black-Scholes volatility. We also demonstrate how to calculate realistic implied volatility surfaces by assuming parameterisations for fractional volatility and the Hurst exponent. Sections 2.3 and 2.4 demonstrate how arbitrage-free calibration is conducted. An empirical analysis using South African equity index and currency option data is presented in Section 2.5. This includes calculating implied Hurst exponents. Finally, Section 2.6 concludes and outlines some ideas for further research.

2.2 Implied Volatility in a Fractional Black-Scholes Market

Since the early 1970s, option pricing has been characterised by the seminal Black–Scholes option pricing formula, which gives a simple bijective mapping between an option's price and the formula's volatility parameter σ_{BS}, termed the option's *implied volatility*. Under the idealised, theoretical assumptions of the Black–Scholes framework, implied volatility is constant. However, when implied volatility is plotted against option strikes for a fixed expiry, one observes a 'skew' or 'smile' pattern in practice, largely driven by the non-normality of the underlying asset return distribution and the supply-demand dynamics within the selected derivatives market (Dupire, 2006). Furthermore, when implied volatility is plotted against option term for a fixed strike, one observes a non-constant relationship, referred to as the term structure of implied volatility.

In reality, then, implied volatility is a function of an option's strike and term. The practitioner's convention in derivatives markets is to speak of separate implied volatility skews (or smiles) for individual option expiries. A collection of implied volatility skews is referred to as an implied volatility surface, which in itself is dynamic, changing with the underlying market conditions (Cont & da Fonseca, 2002). The implied volatility surface at time t is thus denoted as $\sigma_{BS}(K, \tau, t)$.

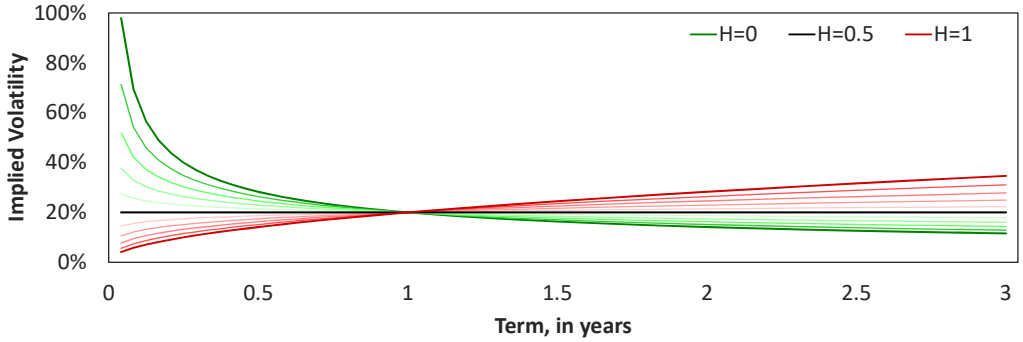


Figure 2.1. Possible implied volatility term structures in a fractional Black– Scholes market for different Hurst exponents and $\sigma_f = 20\%$.

Hu and Øksendal (2003) showed that the variance of the log returns of the stock process in a fractional Black–Scholes market is given by

$$\mathbb{V}ar\left[\ln\left(\frac{S_{t+\tau}}{S_t}\right)\right] = \sigma_f^2 \tau^{2H}, \tag{2.8}$$

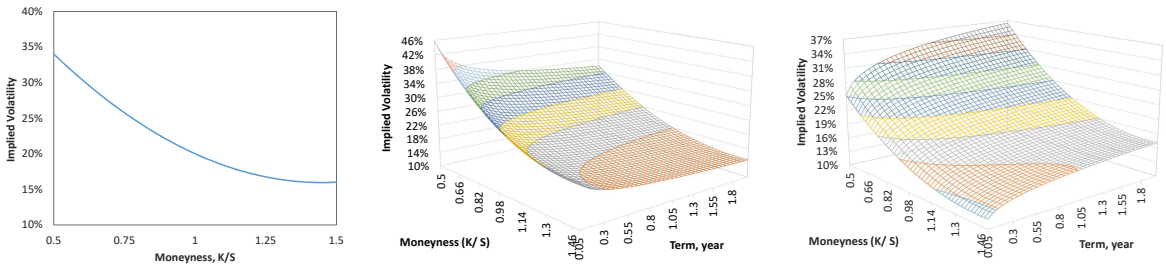
where σ_f is the volatility parameter specific to the fractional setting, hereafter referred to as *fractional volatility*. Equating this expression with the equivalent formula in the standard Black–Scholes market (i.e. substituting in $H = \frac{1}{2}$ above) and dropping strike- and time-dependence for now yields the relationship

$$\sigma_{BS}(\tau) = \sigma_f \tau^{H-\frac{1}{2}}. \tag{2.9}$$

Equation 2.9 has three clear implications. Firstly, even for constant fractional volatility and Hurst exponent, the Black–Scholes implied volatility term structure is described by a power function rather than a constant. This is the same functional form used in Heston's (1993) stochastic volatility model and is also the deterministic term structure function postulated by many market practitioners (Gatheral, 2011). As shown in Figure 2.1, $H > \frac{1}{2}$ gives an up-sloping term structure, $H = \frac{1}{2}$ gives a constant value and $H < \frac{1}{2}$ gives a downward-sloping term structure.

Secondly, the standard and fractional Black–Scholes models give the same implied volatility – and thus option price – for $\tau = 1$, regardless of the specified Hurst exponent. This is also evident from Figure 2.1. It follows that if one assumes constant fractional model parameters, then it must be that $\sigma_f = \sigma_{BS}(1)$.

Thirdly, there is no implicit strike dependence in the fractional Black–Scholes model. This means that the single volatility term structure would apply to all option strikes, which is not consistent with reality. At the very least, one would need to introduce strike dependence into the fractional volatility parameter in order to match the $\tau = 1$ implied volatility skew, which is independent of Hurst exponent by construction. The simplest deterministic model used in practice that gives a reasonable description



(a) 1-year implied volatility **(b)** Volatility surface, $H = 0.4$ **(c)** Volatility surface, $H = 0.6$

Figure 2.2. Indicative implied volatility surfaces for a given 1-year volatility skew and for $H = \{0.4, 0.6\}$.

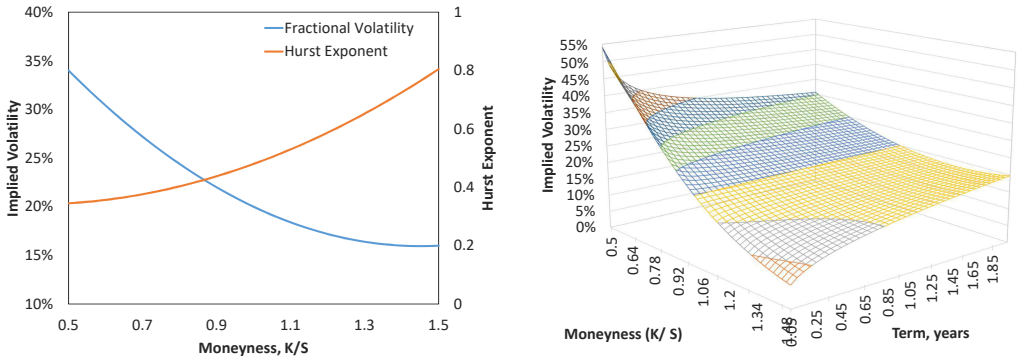
of the implied volatility skew around current spot levels is a quadratic equation (Dumas et al., 1998):[1]

$$\sigma_f = \beta_0 + \beta_1 X + \beta_2 X^2, \tag{2.10}$$

where X is the ratio of the option strike K to the current spot price S_t, generally termed *moneyness*. The β_i parameters account for the level, slope and curvature of the volatility skew, respectively. Figure 2.2 shows how different Hurst exponents can affect the constructed implied volatility surface for a fractional volatility skew (i.e. 1-year implied volatility skew) indicative of equity index option markets.

While the surfaces shown in Figure 2.2 are generally quite realistic, neither captures the universal property that all implied volatility surfaces based on martingale models flatten out with term (Rogers & Tehranchi, 2010). This inconsistency is particularly evident for the $H = 0.6$ surface, which displays increasing skew and curvature across term. In general, for the majority of index volatility surfaces the Hurst exponent would need to be below 0.5 for low strikes and above 0.5 for high strikes to ensure that the surface flattens with term. In contrast, for currency implied volatility surfaces which show considerably more convexity than their equity index counterparts, one would expect the Hurst exponent to be below 0.5 for both very high and very low option strikes. While these expectations stem purely from the mathematics of Equation 2.9 and the shape of volatility surfaces observed in practice, given the stylised facts already known about each asset class, it would seem plausible to assume that there is an underlying economic rationale to the strike profile of the Hurst exponent. This point will be revisited later but for now, we simply observe that realistic index and currency volatility surfaces would require a strike-dependent Hurst exponent.

[1] Even though a quadratic volatility function does not satisfy Lee's moment formula (Roper, 2010) and therefore is not arbitrage free across all strikes, it is still widely used in practice (Tompkins, 2001; Kotzé & Joseph, 2009; Kotzé et al., 2013).



(a) Fractional volatility and Hurst exponent **(b)** Implied volatility surface

Figure 2.3. Indicative implied volatility surface for fractional volatility and Hurst exponent modelled as a quadratic function of strike.

Figure 2.3 shows the implied volatility surface constructed when using a similar deterministic quadratic function as per Equation 2.10 for the Hurst exponent. Notice the significant level of skew achieved at very short option terms – a feat which many stochastic volatility models struggle to achieve (Gatheral, 2011) – in combination with a substantially flatter surface at longer terms.

While the quadratic formulations used here are purely for pedagogical purposes, it is evident that even these simple parameterisations provide one with a high degree of flexibility for modelling realistic implied volatility surfaces in the fractional Black–Scholes framework. Moreover, the idea of using strike-dependent fractional parameters in Equation 2.9 provides one with the basis for a robust but simple implied volatility surface model.

2.3 Arbitrage-free Fractional Black-Scholes Inspired Volatility Surfaces

Creating arbitrage-free parameterisations of the implied volatility surface is extremely important for derivatives trading and risk management in practice and has been given considerable attention in the literature (Lee, 2004; Roper, 2010; Damghani & Kos, 2013; Gatheral & Jacquier, 2014, and the references therein). In this section, we consider a fractional Black-Scholes inspired (FBSI) parameterisation of the volatility surface: a combination of the fBm framework outlined in Section 2.2 and the stochastic volatility inspired (SVI) model of Gatheral (2004) for the strike-dependent fractional volatility parameter.

Carr et al. (2005) introduced the idea of static arbitrage, and Carr and Madan (2005) identified the sufficient conditions – eliminating call spread, butterfly spread and

calendar spread arbitrages – for ensuring that a set of option prices excludes all static arbitrage. Roper (2010) extended this line of research to find the corresponding set of necessary and sufficient conditions to ensure that the volatility surface was free from all static arbitrages. Following the notation of Gatheral and Jacquier (2014), we outline these conditions – no calendar spread arbitrage and no butterfly spread arbitrage – below.

Let us define $k = \ln\left(\frac{K}{F}\right)$ as the log moneyness measured relative to the forward F and $w(k,\tau) = \tau\sigma_{BS}^2(k,\tau)$ as the total implied variance surface. Then, assuming that dividends are proportional to the underlying asset price, the total variance surface w is free from calendar spread arbitrage if and only if

$$\partial_\tau w(k,\tau) \geq 0, \quad \forall\, k \in \mathbb{R},\, \tau > 0. \tag{2.11}$$

Furthermore, each time slice of the total variance surface $w(k)$ is free from butterfly spread arbitrage if and only if the corresponding density function is non-negative, or equivalently

$$g(k) := \left(1 - \frac{kw'(k)}{2w(k)}\right)^2 - \frac{w'(k)^2}{4}\left(\frac{1}{w(k)} + \frac{1}{4}\right) + \frac{w''(k)}{2} \geq 0, \quad \forall\, k \in \mathbb{R}, \tag{2.12}$$

and

$$\lim_{k\to\infty} d_+(k) = \lim_{k\to\infty}\left(\frac{-k}{\sqrt{w(k)}} + \frac{\sqrt{w(k)}}{2}\right) = -\infty. \tag{2.13}$$

Note that $w'(k)$ and $w''(k)$ refer to the first and second derivatives, respectively. Damghani and Kos (2013) give a necessary but not sufficient butterfly spread condition which they state is commonly used in practice:

$$|\partial_k w(k,\tau)| \leq 4, \quad \forall k \in \mathbb{R},\, \tau > 0. \tag{2.14}$$

Let us now consider the fractional Black–Scholes framework as per Section 2.2. It follows from Equation 2.9 that the total implied variance surface at a given time can be written as

$$\begin{aligned} w(k,\tau) &= \sigma_f^2(k)\,\tau^{2H(k)} \\ &= v_f(k)\,\tau^{2H(k)}, \end{aligned} \tag{2.15}$$

where the formulations for fractional variance $v_f = \sigma_f^2$ and Hurst exponent remains fully general. Applying the condition in Equation 2.11, we have that Equation 2.15 is free from calendar spread arbitrage if and only if

$$2v_f(k)\,H(k)\,\tau^{2H(k)-1} \geq 0, \quad \forall k \in \mathbb{R}, \tau > 0. \tag{2.16}$$

Given that $H \in [0,1]$ by construction and $\nu_f > 0$, it is trivial to see that Equation 2.16 will hold true at all times. Therefore, regardless of the parameterisations specified

for fractional volatility and Hurst exponent, the fractional Black–Scholes volatility surface is always free from calendar spread arbitrage. The same conclusion cannot be easily discerned for butterfly spread arbitrage.

As mentioned above, we limit our focus to Gatheral's (2004) SVI model as a candidate for the fractional variance function. The SVI model is one of the most widely used deterministic volatility functions in the equity derivatives market and is also commonly used by foreign exchange derivatives practitioners. Although Gatheral and Jacquier (2014) have recently proposed several alternative formulations of the model parameters, we consider the original 'raw' parameterisation for simplicity. For a given parameter set $\chi = \{a, b, \rho, m, \sigma\}$, the SVI model for total implied variance is given by

$$w(k;\chi) = a + b\left\{\rho(k-m) + \sqrt{(k-m)^2 + \sigma^2}\right\}, \tag{2.17}$$

where $a \in \mathbb{R}$ gives the overall level of variance, $b \geq 0$ gives the angle between the left and right asymptotes, $|\rho| < 1$ determines the orientation of the curve, $m \in \mathbb{R}$ controls the horizontal positioning of the curve and $\sigma > 0$ adjusts the smoothness of the curve vertex. Gatheral (2004) also imposes the condition that $a + b\sigma\sqrt{1-\rho^2} \geq 0$ in order to ensure that $w(k;\chi) \geq 0$ for all $k \in \mathbb{R}$. Gatheral further states that in order to meet the necessary (but not sufficient) condition for no butterfly arbitrage as per Equation 2.14, one must have

$$b(1 + |\rho|) \leq \frac{4}{\tau}. \tag{2.18}$$

Although Roper (2010) showed that a parameter set which satisfies Equation 2.18 can still breach the more stringent Equation 2.12 and thus admit butterfly arbitrage, Gatheral (2004), among others, suggests that the SVI parameter sets calibrated to real market data are arbitrage-free.

As noted in Section 2.2, fractional variance is equivalent to 1-year total implied variance and is thus independent of the Hurst exponent. Therefore, one can directly apply Equations 2.17 and 2.18 in order to find the necessary arbitrage-free SVI parameter ranges. Specifically, for the $\tau = 1$ fractional variance time slice, the necessary condition for no butterfly arbitrage is $0 \leq b \leq \frac{4}{1+|\rho|}$.

Similarly ensuring no arbitrage across all volatility time slices is not easy because of the strike-dependent Hurst exponent. Taking the derivative with respect to strike of the total variance surface as per Equation 2.15, we have

$$|\partial_k w(k,\tau)| = \left|\tau^{2H(k)}\left[v'(k) + v(k)\ln(\tau)\,2H'(k)\right]\right| \leq \frac{4}{\tau}. \tag{2.19}$$

Even for simple $H(k)$ functions, it is not obvious what the necessary arbitrage-free parameter ranges should be. However, it is a straightforward, if somewhat long-winded, exercise to directly calculate the values of $g(k)$ for a given Hurst parameterisation and thus enforce the necessary Hurst exponent ranges during calibration to remove any butterfly spread arbitrage.

2.4 Calibrating FBSI Volatility Surfaces and Implied Hurst Exponents

Building from Sections 2.2 and 2.3, we formally define the fractional Black-Scholes inspired, or FBSI, parameterisation of total implied variance as follows:

$$\begin{aligned} w(k,\tau) &= v_f(k)\,\tau^{2H(k)} \\ v_f(k) &= a + b\left\{\rho(k-m) + \sqrt{(k-m)^2+\sigma^2}\right\} \\ H(k) &= \beta_0 + \beta_1 X + \beta_2 X^2. \end{aligned} \tag{2.20}$$

Motivated by the observations in Sections 2.2 and 2.3, and in the absence of prior knowledge, the choice of a quadratic function for the Hurst exponent seems a reasonable guess. In this case, $\beta_0 \in [0,1]$ represents the at-the-money (ATM) level, β_1 the slope and β_2 the curvature of the Hurst exponent.[2] The function $g(k)$ can be calculated analytically from Equation 2.12 and used to ensure that, in conjunction with the SVI parameter bounds given in Section 2.2, the calibrated β_i parameters do not introduce butterfly arbitrage at any time slice. The complete volatility surface is thus a function of eight parameters, $\chi_f = \{a, b, \rho, m, \sigma, \beta_0, \beta_1, \beta_2\}$.

Given the reliance on the SVI model to parameterise the fractional variance, it makes sense to augment existing SVI calibration algorithms for the additional Hurst exponent parameters. De Marco and Martini (2009) outline a robust quasi-explicit calibration process for the SVI model which produces a reliable and stable parameter set. Through a clever change of variables, the initial five-dimensional SVI minimisation problem is recast into a much simpler two-dimensional problem, with the remaining three variables having (quasi-)explicit solutions within the new framework. See Appendix A.1 for more detail. This '2+3' procedure is robust to initial guesses and provides stable, arbitrage-free SVI parameters.

In a similar vein, we reformulate the raw eight-parameter FBSI model calibration into a '5+3' procedure, with the three Hurst exponent parameters supplementing the two SVI parameters as per De Marco and Martini (2009). Testing shows that this

[2] At-the-money, or ATM, refers to when the option strike is equal to the underlying asset price, which in this case is the forward, i.e. $k = 0$.

procedure is also generally robust to initial guesses and fast to implement. The FBSI model and its calibration procedure thus gives one a robust means of modelling the full implied volatility surface as well as the implied Hurst exponent across the full moneyness range.

To the authors' best knowledge, the only other research to date that considers similar fBm-based volatility surface parameterisations is the fBm variance term structure model posited by Li and Chen (2014).[3] Based on the relationship between implied volatility in the Black–Scholes framework and implied volatility in the fBm framework, Li and Chen (2014) show that one can estimate both the fractional volatility and the implied Hurst exponent from traded option data via linear regression. Consider the logarithm of the power function given in Equation 2.9:

$$\ln\left[\sigma_{BS}\left(\tau\right)\right] = \ln\left(\sigma_f\right) + \left(H - \frac{1}{2}\right)\ln\left(\tau\right). \tag{2.21}$$

Li and Chen (2014) suggest using ordinary least squares (OLS) to estimate the fractional volatility and implied Hurst exponent by regressing the logarithm of ATM implied volatility against the logarithm of term. In this way, one is able to calculate a single fractional volatility and Hurst exponent from the option data.

Li and Chen (2014) further suggest replacing the Black–Scholes implied volatilities in Equation 2.21 with the model-free implied volatilities of Britten-Jones and Neuberger (2000), which can be calculated in practice by applying the standard VIX methodology at all observed option terms.[4] The use of model-free implied volatility as the dependent variable has the benefits of removing dependence on any specific pricing model and of using information from all traded options rather than only ATM options.[5] However, despite incorporating information from the full volatility surface, this method still only allows one to model the term structure of implied volatility.

[3] Although fractional volatility models have been around since the work of Comte and Renault (1996, 1998) and Baillie et al. (1996), this is essentially a subfield of the much larger stochastic volatility literature, where fractional noise rather than Gaussian noise is used within the volatility process. In comparison, this work differs in three aspects. Firstly, the use of fractional noise is restricted to the stock price process. Secondly, this work falls within the deterministic rather than stochastic volatility modelling literature. Thirdly, the Hurst exponent is assumed to be a non-constant function of strike and time rather than a constant parameter in a volatility process.

[4] VIX refers to the Chicago Board Options Exchange Volatility Index®, a volatility benchmark index based on options on the S&P 500® Index.

[5] This second benefit stems from the fact that model-free implied volatility is calculated using the complete volatility skew at each term.

2.5 Empirical FBSI surfaces and Hurst Exponents: A South African Experiment

The FBSI and Li and Chen (LC) models are calibrated to two sets of South African option market data. The first data set consists of 529 weekly observations of implied volatility skews for listed futures options on the FTSE/JSE Top40 index (Top40) over the period 5 September 2005 to 30 November 2015. Top40 options are the most actively traded option contracts in South Africa. These options trade on the Johannesburg Stock Exchange (JSE) Derivatives Exchange on the basis of implied volatility and the option price is calculated using the Black (1976) option pricing formula. Weekly implied volatility skew observations were obtained from Legae Peresec and cover a strike range of $75\% - 125\%$ of the forward price. The second data set, also obtained from Legae Peresec, consists of 146 weekly observations of implied volatility skews for listed futures options on the United States Dollar to South African Rand (USDZAR) exchange rate over the period 11 February 2013 to 30 November 2015. The implied volatility skews up to November 2014 cover a range of $80\% - 120\%$ of the forward price and thereafter cover a $70\% - 130\%$ range.

2.5.1 FBSI Equity Index Volatility Surfaces

Let us first consider the results for the index volatility surfaces. Figure 2.4 displays Top40 index performance since September 2005 against the fractional volatility and implied Hurst exponent from the calibrated FBSI volatility surface model and the LC volatility term structure model.

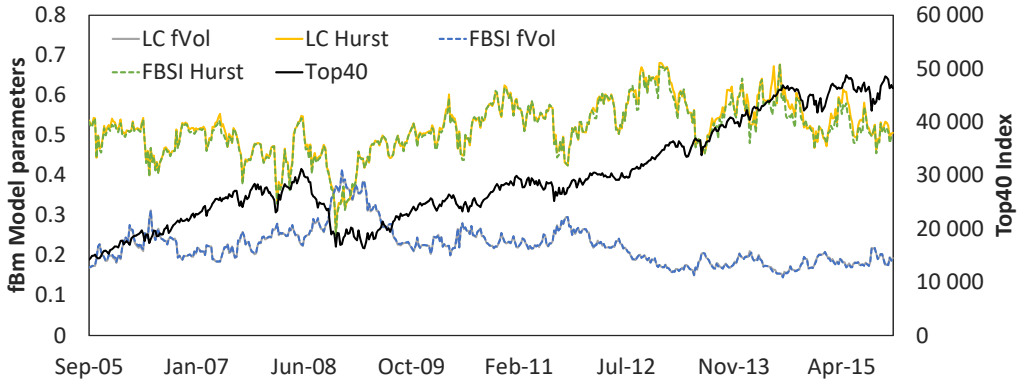


Figure 2.4. Top40 index performance versus ATM fractional volatility and implied Hurst exponent from the calibrated FBSI and LC models, Sep 2005 to Nov 2015.

Visual inspection confirms the well-documented inverse relationship between index performance and fractional volatility (i.e. 1-year implied volatility) and also suggests a positive relationship between index performance and the implied Hurst exponent,

particularly evident during the 2008 financial crisis. This is confirmed by the moderately positive correlation values of 0.47 and 0.45 displayed in Table 2.1 between Top40 log returns and the log returns of each model's implied Hurst exponent series.

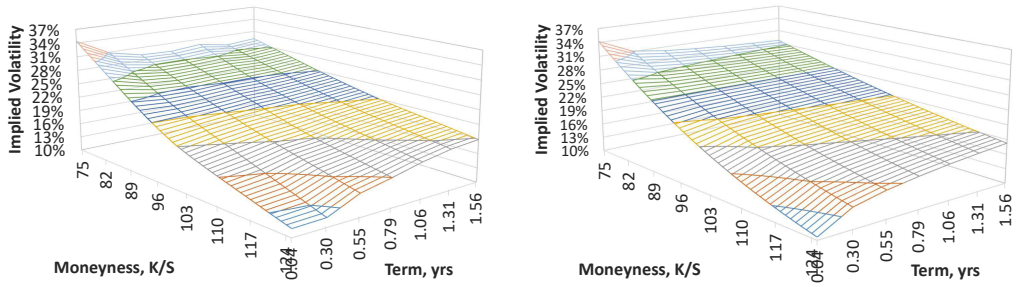
Table 2.1. Correlation matrix of weekly log returns on Top40 index and associated implied volatility parameters, Sep 2005 to Nov 2015.

	Top40	**LC σ_f**	**LC Hurst**	**FBSI σ_f**	**FBSI Hurst**
Top40	1				
LC σ_f	-0.516	1			
LC Hurst	0.473	-0.338	1		
FBSI σ_f	-0.514	0.992	-0.339	1	
FBSI Hurst	0.448	-0.313	0.956	-0.284	1

However, there are also times when one sees significant changes in the implied Hurst exponent without any large associated downturns in the index. For example, the Hurst exponent fell materially from a high of 0.67 down to 0.46 during the first half of 2013, while the index remained range-bound around the 35 000 level. Over the same period, fractional volatility also remained fairly stable between 16% and 18% and only picked up briefly around the middle of 2013. This suggests that fractional volatility and implied Hurst exponent capture somewhat different aspects of the total uncertainty within the index and thus provide one with more detailed information on the underlying price process.

This suggestion is borne out by the correlation seen between the log returns in fractional volatility and implied Hurst exponent, shown in Table 2.1. Although negative as one would expect, it is considerably lower in absolute terms than the correlations displayed between the respective parameters and the underlying index returns. Therefore, deconstructing the single implied volatility number into a long memory component and a long memory-conditioned volatility component may well have useful application in a wide range of financial applications, including derivatives trading, risk management and dynamic asset allocation. For example, discrete delta-hedging strategies could potentially be improved by incorporating the implied Hurst exponent as a means of identifying how rough or smooth the index returns are likely to be and also whether the index is currently more likely to mean-revert or continue trending. For now, we leave such applications of the implied Hurst exponent for future research.

Notice that the ATM FBSI fractional volatility time series is nearly identical to the LC fractional volatility series, with a correlation of 0.99. The ATM Hurst exponent time series is also very similar across models with a correlation of 0.96, although slight

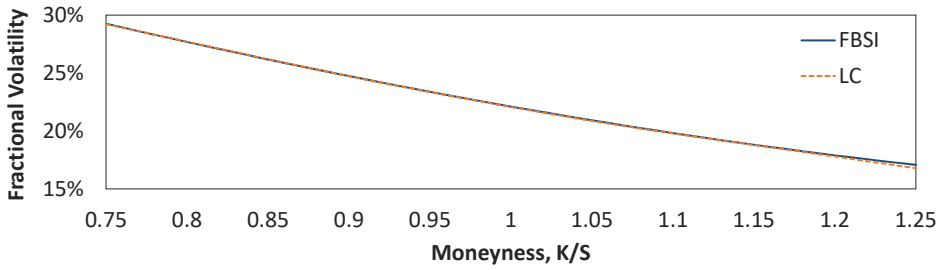


(a) Actual Top40 volatility surface

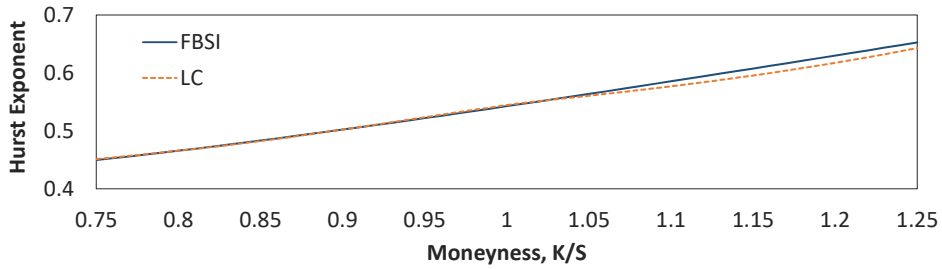
(b) Calibrated FBSI volatility surface

Figure 2.5. Observed Top40 implied volatility surface versus calibrated FBSI surface as at 30 May 2011.

deviations are evident in the final 2 years of the sample period. This high degree of equivalence indicates that the FBSI model provides sufficient flexibility to model the ATM term structure accurately even while fitting the complete index volatility surface. Figure 2.5 confirms this by displaying the Top40 traded volatility surface and its calibrated FBSI counterpart as at 30 May 2011. The modelled surface mirrors the market surface very well at most terms and moneyness levels, although there are some small areas on the market surface where the power law term structure assumption is violated.



(a) FBSI versus LC fractional volatility



(b) FBSI versus LC implied Hurst exponent

Figure 2.6. Calibrated Top40 fractional volatility and implied Hurst exponent skews from the FBSI and LC models as at 30 May 2011.

The reason why the FBSI model fits the equity surfaces so well is shown in Figure 2.6. The calibrated FBSI parameter curves are compared to those obtained from

separately fitting the LC term structure model for each moneyness level. For our data, this equates to running 51 independent regressions, which ensures a very accurate fit of the surface thanks to the use of 102 parameters. Although clearly not a viable candidate for modelling the surface directly, this LC 'multi-model' provides one with an excellent means of evaluating whether the quadratic and SVI functions provide sufficient flexibility for capturing the required strike dependence in fBm volatility parameters.

As Figure 2.6 shows, the fractional volatility curves from both models are essentially equivalent, while the FBSI Hurst exponent shows a slight deviation from the LC multi-model curve above the 105% moneyness level. This discrepancy is responsible for the difference at high moneyness levels and very short terms between the traded and fitted volatility surfaces shown in Figure 2.5.

2.5.2 FBSI Currency Volatility Surfaces

Figure 2.7 shows the FBSI and LC model parameters from February 2013 in comparison with the underlying USDZAR foreign exchange rate. In contrast to the equity index results given in Section 2.5.1, there are significant differences between the FBSI and LC implied Hurst exponents evident across the full sample period. The FBSI implied Hurst exponent is almost always lower than its LC counterpart and the positive correlation of 0.36 is much lower than one would expect given that both time series represent the same parameter. Fractional volatility is far more similar across the two models, with a correlation of 0.83. There are still noticeable differences though, with FBSI fractional volatility also generally lower than LC fractional volatility across the period.

Table 2.2 also shows the expected positive relationship between exchange rate and fractional volatility, evident in both the LC ($\rho = 0.44$) and FBSI ($\rho = 0.45$) models. Interestingly though, the negative relationship evident between the exchange rate and the LC implied Hurst exponent ($\rho = -0.44$) is considerably stronger than that between the exchange rate and the FBSI implied Hurst exponent ($\rho = -0.16$). From Figure 2.7 we see that although the USDZAR has consistently trended upwards over the sample period, both implied Hurst exponent and fractional volatility parameters remained largely range-bound for most of the period. Only over the last year has one seen a slight decline in implied Hurst exponent levels and a concurrent increase in fractional volatility levels as the size of the weekly exchange rate moves has grown.

Table 2.2 also shows that the correlation between LC parameters is weak and negative, while that between the comparative FBSI parameters is instead weak

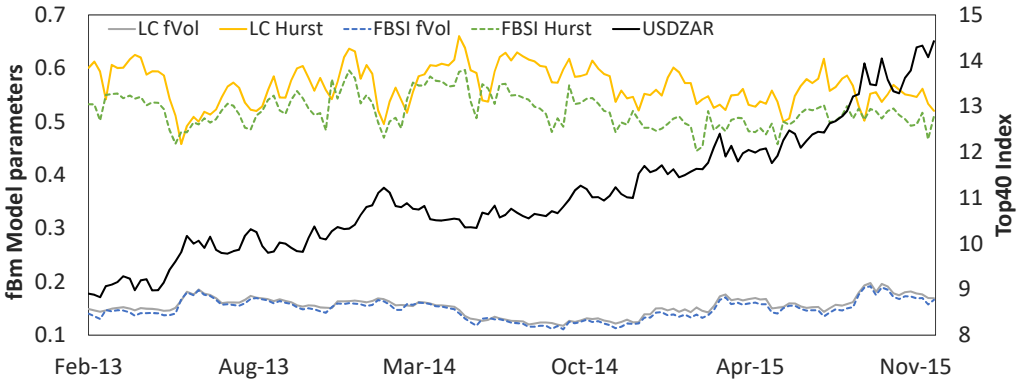


Figure 2.7. USDZAR performance versus ATM fractional volatility and implied Hurst exponent from the calibrated FBSI and LC models, Feb 2013 to Nov 2015.

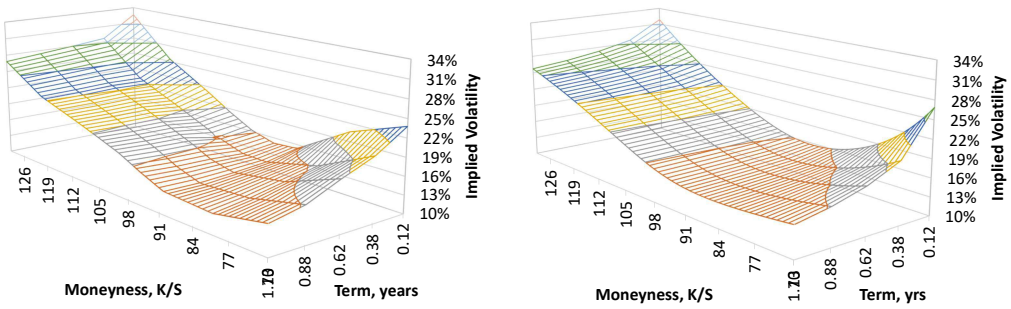
and positive. This again suggests a certain level of independence between the two implied volatility components.

The large differences between the FBSI and LC parameters indicate that, in its current form, the FBSI model is unable to adequately replicate the currency implied volatility surface. Figures 2.8 and 2.9 show the problem for an example currency surface as at 28 April 2015. The traded volatility skews are significantly sloped for strikes above the forward level (i.e. high moneyness) and remain so even for longer terms. In contrast, the trade skews are less sloped for strikes below the forward level (i.e. low moneyness) and flatten off a fair degree with term. However, because the traded short-term volatility skew flattens out at the lowest moneyness levels, this means that the curvature of the term structures will also flatten out at the lowest moneyness levels.

Table 2.2. Correlation matrix of weekly log returns on USDZAR and associated implied volatility parameters, Feb 2013 to Nov 2015.

	USDZAR	**LC σ_f**	**LC Hurst**	**FBSI σ_f**	**FBSI Hurst**
Top40	1				
LC σ_f	0.438	1			
LC Hurst	0.442	-0.175	1		
FBSI σ_f	0.451	0.834	-0.360	1	
FBSI Hurst	-0.159	-0.149	0.359	0.240	1

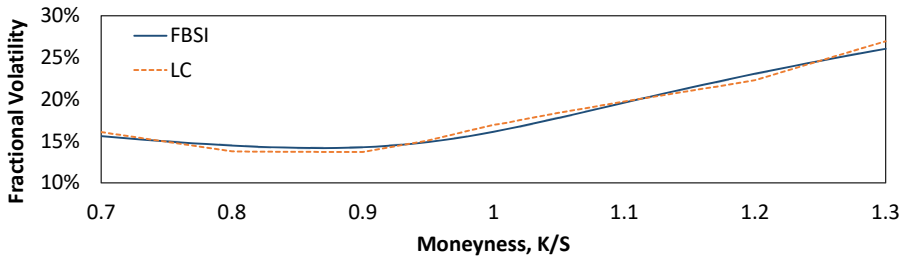
These observations together suggest that the implied Hurst exponent would not only need to be generally concave but also include inflection points at low moneyness levels and possibly also at high moneyness levels, as shown in the lower panel of Figure 2.9. The assumed quadratic function is not capable of this, and thus the calibrated Hurst exponent function represents a trade-off between matching the required level of ATM convexity and minimising the mismatch for far out of the money volatility points. Therefore, we would suggest using a different functional



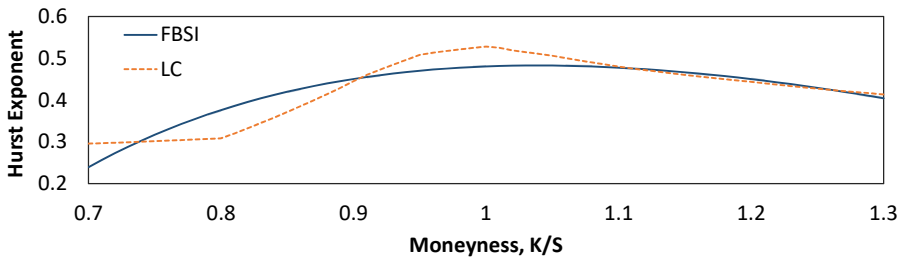
(a) Actual USDZAR volatility surface **(b)** Calibrated FBSI volatility surface

Figure 2.8. Observed USDZAR implied volatility surface versus calibrated FBSI surface as at 28 April 2015.

form for the implied Hurst exponent in the currency derivatives space. Given the need for an inflection point in the Hurst exponent curve, the most obvious starting point would be a third-order polynomial. For now though, we leave this remark as an avenue for future research. The calibrated FBSI volatility surface shown in Figure 2.8 still manages to capture most of the traded surface's characteristics with the added benefit of being fully analytic; an important consideration when valuing exotic derivatives under local volatility.[6]



(a) FBSI versus LC fractional volatility



(b) FBSI versus LC implied Hurst exponent

Figure 2.9. Calibrated USDZAR fractional volatility and implied Hurst exponent skews from the FBSI and LC models as at 28 April 2015.

[6]Local volatility, introduced by Dupire (1994) and Derman et al. (1996), generalises the Black-Scholes model by treating asset volatility as a deterministic function of asset price and time. It is commonly used to value exotic options in a manner that is consistent with observed vanilla option prices.

2.6 Conclusion

This chapter addresses several theoretical and practical issues in option pricing and implied volatility calibration in a fractional Black–Scholes market. We start off by discussing how options can be priced when the noise component of the underlying risky asset is driven by a fractional Brownian motion. We then describe the links between standard Black–Scholes volatility and fractional Black–Scholes volatility and highlight two important observations. Firstly, the fractional Black–Scholes model admits a non-constant implied volatility term structure when the Hurst exponent is not equal to 0.5. More specifically, this term structure is described by a power function and is up-sloping (down-sloping) when the Hurst exponent is greater (less) than 0.5. Secondly, 1-year implied volatility is independent of the Hurst exponent and equivalent to fractional volatility.

Building on these two observations, we show how one can construct realistic implied volatility surfaces by assuming simple parameterisations for fractional volatility and Hurst exponent. In particular, we introduce the eight-parameter FBSI model. This novel deterministic volatility surface model is based on the fractional Black–Scholes framework and uses Gatheral's (2004) SVI parameterisation for the fractional volatility skew and a quadratic parameterisation for the implied Hurst exponent skew. One benefit of this model is that it provides us with a parsimonious decomposition of the implied volatility surface into an independent long memory component and a conditional volatility component. Such a decomposition could be usefully applied in a wide range of financial applications, including derivatives trading, risk management and dynamic asset allocation.

We address the issue of arbitrage-free calibration for the FBSI model in depth and prove in general that any FBSI volatility surface will be free from calendar spread arbitrage. Although one cannot make a similar statement about butterfly spread arbitrage, we show that it is simple to control for this during the calibration process because of the fully analytical form of the surface.

Finally, we test the FBSI model empirically against Li and Chen's (2014) volatility term structure model using implied volatility surfaces on South African listed Top40 index futures options and listed USDZAR currency futures options. We find that the FBSI model fits the equity implied volatility surfaces very well and, furthermore, that the decomposition of implied volatility into its long memory and fractional volatility components provides one with more detailed information on the true uncertainty within the underlying asset price process. The currency implied volatility surfaces provide more of a calibration challenge for the FBSI model because of a flattening in the term structure at far out of the money strikes. The calibrated FBSI volatility

surface still manages to capture most of the traded surfaces' characteristics though with the added benefit of being fully analytic; an important consideration when valuing exotic derivatives under local volatility.

There exist several avenues for further research based on the work in this chapter, of which we will highlight two. Firstly, researchers could consider alternative functional forms for the implied Hurst exponent in order to better model the more complex currency option volatility surface. Secondly, researchers could examine the use of the implied Hurst exponent in a number of trading and risk management settings. One example would be to explore the use of the implied Hurst exponent in a delta hedging strategy as an adjustment factor accounting for the forward-looking persistence or smoothness of the underlying asset process.

3

Implied Distributions and Recovery in Illiquid Markets

> *My sources are unreliable, but their information is fascinating.*
>
> — **Ashleigh Brilliant**
> (Author and cartoonist)

Chapter Synopsis

In this chapter, we describe how forward-looking information on the statistical properties of an asset can be extracted directly from options market data and how this can be used practically in portfolio management.

Although the extraction of a forward-looking risk-neutral distribution is well-established in the literature, the issue of estimation in an illiquid market is not. We use the deterministic SVI volatility model to estimate weekly risk-neutral distribution surfaces. The issue of calibration with sparse and noisy data is considered at length and a simple but robust fitting algorithm is proposed.

Furthermore, we extract real-world implied information solely from options data by implementing the recovery theorem introduced by Ross (2015). Recovery is an ill-posed problem that requires careful consideration. In this research, we describe a regularisation methodology for extracting real-world implied distributions and implement this method on a history of equity index SVI implied volatility surfaces. We analyse the first four moments from the implied risk-neutral and real-world implied distributions and use them as signals within a simple tactical asset allocation framework, finding promising results.

This chapter is adapted from the journal article by Flint and Maré (2017a) and addresses research questions 1b, 2a and 3a given in Section 1.2.3.

3.1 Introduction

An important requirement for optimal portfolio construction is an understanding of the future possible returns of the constituent assets. Armed with this understanding, investors can ensure that their chosen combination of assets will lead to a portfolio that is consistent with their risk tolerances and return objectives. Unfortunately, forecasting return distributions accurately is a challenging endeavour (Cochrane, 2011). A common approach is to use historical data as the basis for forecasts. For example, standard deviation and expected returns are easily estimated from historical data and, when combined with an assumption of normally distributed returns, provide a completely specified return distribution. Unfortunately, empirical studies have shown that expected return and standard deviation values estimated from historical data are unstable and the assumption that historical estimates will apply at a future date corresponding to the investment horizon is questionable at best (see, e.g. Michaud, 1989). An alternate forecasting method is to extract forward-looking information on the statistical properties of an asset directly from options market data.

The seminal work of Black and Scholes (1973) and Merton (1973b) proved that the value of an option in a complete market is independent of the expected return on the underlying asset and thus gave rise to the risk-neutral valuation framework. Under this framework, the only unknown parameter affecting an option's value is the volatility of the underlying asset, referred to as the implied volatility. Because of this, the Black–Scholes–Merton (BSM) pricing formula has become ubiquitous in derivatives markets worldwide due to its ability to monotonically translate any option price into a single, easily-comparable implied volatility value. In this sense, implied volatility is the one language common to all option markets.

Implied volatility surfaces in practice differ in three important ways from the flat theoretical BSM surface. Firstly, implied volatility varies with the strike of an option. Secondly, implied volatility varies with option term. Thirdly, the shape of the volatility surface changes over time depending on the underlying market regime and trading conditions. It is generally accepted that the volatility surface represents a combination of the consensus view of the terminal asset return distribution, current market risk preferences and any supply-demand factors stemming from structural market issues. Therefore, in addition to providing one with a means of pricing options, the implied volatility surface can be viewed as containing the sum of all forward-looking information known (or assumed) about the underlying asset.

The idea of accessing this embedded information is not new. Implied volatility has long been used as a gauge of investor risk sentiment or fear, with the Chicago Board

of Options Exchange Volatility Index (VIX) being the most commonly referenced measure today. However, since the mid 1990s, options have increasingly become assets in their own right and there has been a concerted effort to study the extent of the predictive information embedded in these markets.

Breeden and Litzenberger (1978) proved that the forward-looking risk-neutral distribution (RND) could be extracted directly from an arbitrage-free derivatives market provided that one knows the European option prices for all levels of the underlying. This result gave investors the means to estimate the forward-looking RND for a given term from the implied volatility skew for that same term. A range of interesting statistical metrics can then be calculated from the RND, including volatility (i.e. the VIX), skewness, kurtosis and value-at-risk, all of which are inherently forward-looking by construction. Furthermore, if volatility skews on an equity index and its underlying stocks are available, then it is also possible to estimate forward-looking implied stock correlations and betas. Kempf et al. (2015), Baule et al. (2015), DeMiguel et al. (2013), Buss and Vilkov (2012) and Kostakis et al. (2011), among others, have shown that such implied moments and statistics significantly outperform their historical counterparts in a range of portfolio, risk management and trading applications.

An important point to remember though is that risk-neutral probabilities are not equivalent to real-world probabilities. A change in risk-neutral probabilities can either stem from a change in the underlying real-world probabilities or from a change in underlying risk preferences (Malz, 2014). Furthermore, RNDs do not give one any forward-looking information on the real-world expected return. In fact, until very recently it was considered impossible to extract any real-world information directly from option markets without either having to make stringent assumptions about investors' risk preferences or resorting to estimation of the same from historical data. However, a recent development has brought this belief into contention.

Ross (2015) postulated the recovery theorem, which, for a given set of market and risk preference restrictions, makes it possible to estimate real-world information directly from an options market. Although some of the assumptions underlying Ross's recovery theorem are obvious simplifications of actual market conditions, the more pertinent practical question – as Audrino et al. (2015) point out – is whether this recovered real-world distribution provides any additional insight over that found in its more easily available risk-neutral counterpart. While it remains a fundamental question as to whether there can actually be either unique or better information within a secondary market, the fact of the matter is that option-implied information is currently being used in a wide range of market applications. Therefore, robust estimation of this information is a critical empirical issue. In this work, we contribute to the research in this field by considering in detail the estimation and application of

risk-neutral and real-world option-implied distributions in an illiquid market setting where data are both sparse and noisy.

The rest of this chapter is organised as follows. Section 3.2 tackles the problem of estimating RNDs by introducing a range of common estimation techniques and discussing their applicability in the context of the illiquid South African options market. The chosen technique is then discussed at length and a practical RND estimation algorithm is presented. Section 3.3 introduces the recovery theorem, discusses some implementation challenges and presents an algorithm for applying the recovery theorem using regularised least squares. Empirical results using South African option data are presented in Section 3.4. General option-implied data applications are discussed, recovered real-world distribution moments are compared to their risk-neutral counterparts and a tactical asset allocation example using implied information is presented. Section 3.5 concludes and outlines some ideas for further research.

3.2 Estimating Option-Implied Distributions in Illiquid Markets

In a complete and arbitrage-free market, Cox and Ross (1976) show that the model-free value of a European call option C_t at time t, with term $\tau = T - t$ and strike K is equal to

$$C_t(K,T) = e^{r\tau} \int_0^\infty (S_T - K)^+ q(S_T)\, ds, \tag{3.1}$$

where r is the risk-free rate, S_T is the terminal price of the underlying and $q(S_T)$ is the terminal risk-neutral distribution of the underlying asset. Taking the second derivative with respect to strike yields the seminal result from Breeden and Litzenberger (1978):

$$e^{r\tau} \frac{\partial^2 C}{\partial K^2} = q(K). \tag{3.2}$$

In theory, one needs a continuum of option prices for a given term in order to calculate the RND. In practice though, only a discrete set of options are actually traded and thus some estimation procedure is required. A wide range of RND estimation techniques have been suggested in the literature, which can be broadly classified by whether they work with Equations 3.1 or 3.2.[1]

The techniques based on Equation 3.1 postulate some distributional form for the RND, which is then evaluated based on an objective function measuring the distance between the estimated and actual option prices. Parametric forms include

[1] Obviously, there are some techniques that cannot easily be shoehorned into this classification scheme but we believe that it still provides a useful means of summarising the most popular techniques.

adding expansion terms to base distributions (Coutant et al., 1998), using complex underlying distributions (Posner & Milevsky, 1998), or using mixtures of lognormal distributions (Melick & Thomas, 1997). Nonparametric forms include the use of entropy-based methods (Buchen & Kelly, 1996) and kernel-based methods (Aït-Sahalia & Lo, 1998).

The estimation techniques based on Equation 3.2 instead postulate some continuous form of the underlying volatility skew which can be used to interpolate between the traded option volatilities – and thus prices – as well as extrapolate outside of the traded strike range. The second derivative of the call prices and hence the RND is then found numerically. These techniques can be divided into three sub-categories. Firstly, one can use curve-fitting techniques such as cubic splines to interpolate between and extrapolate from traded option volatilities (Bliss & Panigirtzoglou, 2002). Secondly, one can fit deterministic volatility models to the traded data (Shimko, 1993; Dumas et al., 1998), and thirdly, one can postulate more complex models for the underlying process in the form of stochastic volatility (Heston, 1993), jump diffusion (Merton, 1976), or a combination of the two (Bates, 2000).

Given this vast array of RND estimation techniques as well as the range of different applications in which these RNDs are used, it is perhaps not surprising that it remains an open question as to which technique – if any – is considered 'optimal'. Of the small number of comparative studies done to date, the only universal conclusion is that estimating RNDs is an ill-posed problem, which can be highly dependent on the estimation technique as well as the available data.[2] This means that the choice of technique should be considered an active decision in the RND estimation process as well as in the larger recovery process.

This becomes even more important in illiquid markets where option data are both sparse and noisy. Because of these two issues, many of the techniques proposed above become unsuitable. Apart from the shape-constrained kernel method of Aït-Sahalia and Duarte (2003), the majority of the distributional-based methods are largely unconstrained and thus will struggle under sparse, noisy data conditions as their inherent flexibility may actually augment estimation error. For example, Cooper (1999) shows that under real-world conditions, noisy option data can lead to large spikes in the RNDs estimated using the mixture of lognormals approach. McManus (1999) notes a similar result for the entropy-based techniques. The same argument can be extended to spline-based techniques, which are heavily dependent on the choice and number of knots. Spline-based techniques also raise the additional question of how to extrapolate implied volatility skews beyond the traded range. Malz (2014) and Bliss and Panigirtzoglou (2002; 2004) suggest assuming flat

[2] See, for example, Aït-Sahalia and Duarte (2003), Cooper (1999), McManus (1999) and Coutant et al. (1998).

volatility – and thus lognormal RND tails – outside of the traded range, whereas Figlewski (2008) instead suggests grafting Generalised Extreme Value distribution tails onto the estimated central portion of the RND. In either case, the researcher is ultimately prespecifying the tail structure of the RND, thus actually making the spline technique somewhat parametric.

There is also another issue that needs to be considered in this research. Successful application of the recovery theorem requires RNDs across a number of terms – i.e. an RND 'surface' – rather than the single-term distributions usually considered in the literature. This means that not only does one have to consider arbitrage constraints across strike but also across term. While the kernel methods of Aït-Sahalia and Duarte (2003) do consider the former, they do not formally make provision for the latter. In contrast, the problem of static arbitrage across both strike and term has been extensively researched in the volatility modelling literature.[3] Therefore, given our requirement of a complete arbitrage-free RND surface, this would suggest choosing either the deterministic or stochastic volatility modelling approach. Popular candidates in each area respectively are Gatheral's (2004) stochastic volatility inspired (SVI) model and Heston's (1993) stochastic volatility model, both of which are used extensively by academics and practitioners worldwide.

Of these two candidates, Gatheral (2011) states that stochastic volatility models fail to capture the dynamics of short-term volatility skews and can also be hard to calibrate in practice. On the other hand, the comprehensive study by Tompkins (2001) suggests that most option markets are well modelled by simpler deterministic functions. Furthermore, deterministic models provide the flexibility to calibrate implied volatility separately across strike and term but also the simplicity to ensure arbitrage-free volatility surfaces with minimal model error. Based on these observations, as well our overarching aim to extract information in as robust and flexible a manner as possible, we choose to model the implied volatility surface – and thus the RND surface – using the SVI model.

3.2.1 The South African Options Market

In this study, we consider fully margined options on FTSE/JSE Top40 Index (Top40) futures, traded on the Johannesburg Stock Exchange (JSE) Derivatives Exchange.[4] These listed options expire quarterly on the third Thursday of March, June, September and December each year. West (2005) provides one of the few studies on the volatility calibration challenges faced within this market. At the time of his study,

[3] For example, see Carr et al. (2003), Carr and Madan (2005) and Roper (2010).

[4] The Top40 Index comprises the largest South African companies based on free float-adjusted market capitalisation.

over the counter (OTC) structures comprised the majority of South African (SA) option activity. However, the subprime crisis in 2008 changed the manner in which investors traded. A renewed interest in regulation and a renewed appreciation of credit risk resulted in a significant increase in the number of exchange-traded contracts and a concomitant decline in OTC volumes. This trend was further aided by the introduction of the JSE CanDo derivatives platform, which essentially gave investors the ability to list, trade and margin any exotic derivative directly with the exchange. The shift towards listed contracts also meant a far larger proportion of the underlying trade data became publicly available.

In the SA market, participants mostly use derivatives as hedging tools, meaning that open interest is concentrated in put options with strikes below current index levels. These hedging structures are usually short-term. Volumes are thus concentrated in the three closest expiries – i.e. up to nine months – and trade in any term longer than 15 months is extremely rare. The size of such hedges can also sometime dwarf all other trades in a given period, leading to an extremely skewed open interest distribution across option strikes. That being said, total option volumes remain small in comparison to other developed markets. On any given day, the number of trades varies significantly and there could even be no trades across any expiry. The traded strike range is also quite narrow and generally spans a maximum range of –20% to +15% of current index levels.

Daily listed Top40 option trade data is freely available from the JSE website back to February 2011. We further sourced option trade data back to September 2005 from Legae Peresec, a large derivatives broker in South Africa. For each option trade, the full data set generally includes trade date and time, futures level, strike, traded volatility, price, option type and volume.[5] However, market participants do have some leeway in terms of what and how to report this information to the exchange and so incomplete records can and do occur.

3.2.2 The Stochastic Volatility Inspired Model

The SVI model was disseminated by Gatheral (2004) and has since arguably become the practitioner's model of choice in the equity derivative space (Damghani & Kos, 2013). It is known to fit equity volatility skews extremely well but is still intuitive and easy to implement. Denoting the futures level as F, the term as τ and the strike as K, we can write the SVI implied variance as

$$\sigma^2(x,\tau) = a + b\left(\rho(x-m) + \sqrt{(x-m)^2 + s^2}\right), \tag{3.3}$$

[5]Note that there is no bid-ask spread information included.

where $x = \ln(\frac{K}{F})$ is the log-moneyness and the parameter set $\{a, b, \rho, m, s\}$ is specific to each expiry. This parameterisation was inspired by the large-term asymptotic behaviour of the Heston stochastic volatility model. In essence, the SVI model fits a hyperbola to implied variance in log-moneyness space. This particular form is chosen because it ensures that variance is linear as $|x| \to \infty$ – a fundamental characteristic of volatility surfaces – while still being convex around the at-the-money (ATM) level. This is intuitive for traders in that the more out-the-money (OTM) an option is, the more volatility convexity the option displays. Each SVI parameter has an intuitive geometric interpretation (Gatheral & Jacquier, 2014):

- a defines the overall level of variance and shifts the skew vertically;
- b defines the angle between the put and call wing variance slopes;
- ρ rotates the variance curve clockwise around the current forward level;
- m shifts the variance curve left or right;
- s defines the amount of ATM variance curvature.

The smart choice of parameterisation coupled with the five degrees of freedom generally ensures an extremely good fit in practice, particularly in the equity index space. Furthermore, because of its characterisation, the SVI model is able to provide decent approximations for deep OTM volatilities and can also produce sufficient ATM curvature at very short terms, a known failing of many stochastic volatility models. Finally, Gatheral and Jacquier (2014) also show that the SVI single-expiry calibration process is also easily coupled with calendar-spread arbitrage checks, which enables straightforward construction of smooth, arbitrage-free implied volatility surfaces.

One drawback of the SVI model though is that the usual least squares minimisation of the implied volatility objective function is very sensitive to initial parameter guesses. Furthermore, the function displays several local minima, which can seriously bias final parameter estimates. De Marco and Martini (2009) addressed this shortcoming by finding a robust quasi-explicit calibration process which produced a reliable and stable parameter set. Through a clever change of variables, the initial five-dimensional SVI minimisation problem is recast into a much simpler two-dimensional problem, with the remaining three variables having quasi-explicit solutions within the new framework. This ingenious '2+3' procedure is much less sensitive to initial guesses and provides stable, arbitrage-free SVI parameters. See Appendix A.1 for more detail. Gatheral and Jacquier (2014) suggest another calibration method based on a similar reparameterisation of their original model.

3.2.3 Constructing SVI Volatility and RND Surfaces

Although international literature on modelling implied volatility is vast, the majority of these studies are not easily applicable to the SA derivatives market due to its illiquid nature and fairly unique trading dynamics. To the authors' knowledge, only West (2005) and Kotzé and Joseph (2009) discuss the issue of calibration in such a market. West (2005) calibrates the SABR (stochastic alpha beta rho) model of Hagan et al. (2002) to options on Top40 futures, while Kotzé and Joseph (2009) do the same but using a quadratic deterministic volatility model.[6] Both studies stress the need for robust and sensible calibration algorithms and put forth several useful suggestions for reaching that goal, which are incorporated below. Be that as it may, creating a robust calibration procedure still requires some "creative decision making" as West puts it. In this context, the SVI volatility and RND surface algorithm given here represents a blend of theoretical best practices and market experience in the presence of severe practical constraints. For a given point in time, we construct implied volatility and RND surfaces as follows:

1. Collate Top40 option trade data for the past seven days. Backfill missing values as required using the given information and the Black (1976) pricing equation adjusted for fully margined options. Discard those records which cannot be completed.

2. Apply a daily exponential time-weighting function ($\lambda = 0.915$) to moderately down-weight older trades and a stepped size-weighting function to significantly up- or down-weight trades falling in prespecified size buckets ($w_i = \{0.1, 0.8, 1, 0.8\}$ for trades of less than 100, 500, 2 000 and 10 000 contracts respectively).

3. In cases of extreme data sparsity, include several OTM skew markers from the previous period's calibration, adjusted for the current ATM volatility level.

4. Calibrate the SVI model separately to each traded expiry using the '2+3' algorithm of De Marco and Martini (2009).

5. Check for calendar-spread arbitrage by examining the total variance plot for any crossed lines. If necessary, recalibrate the SVI parameters from shortest to longest expiration and include a large penalty for crossing with the previous skew, as per Gatheral and Jacquier (2014).

[6] This was the model used by the JSE until July 2017 for mark-to-market and margining purposes.

6. Use the (re)calibrated SVI parameters to create volatility skews across a $20 - 300\%$ range of the prevailing forward prices.

7. Interpolate linearly in total variance space between the calibrated expiries to create monthly volatility skews ranging from $1 - 15$ months (total range dependent on available expiries).

8. Calculate call prices across the full strike range at each term from the interpolated volatility skews and estimate the monthly RNDs numerically using Equation 3.2.

We use this fitting procedure to create weekly arbitrage-free implied volatility and RND surfaces over the period 5 September 2005 to 16 May 2016, giving a total of 559 surface observations. The prevailing interest rate and dividend yield curves are also recorded at the calibration dates.

3.3 Recovering Real-World Implied Distributions

Below, we give a brief outline of the Ross recovery theorem along with its underlying assumptions in a style similar to Spears (2013).[7] We then consider some of the technical difficulties in applying the recovery theorem in practice and present our implementation procedure. Note that the illiquid market data issues highlighted above do not directly affect the recovery process as the only required input is an estimated RND surface.

3.3.1 The Ross Recovery Theorem

Before stating the recovery theorem from Ross (2015), we need to introduce several underlying financial concepts. Assume that the underlying asset S_t can only take on a finite n number of states at time t. The transition probability matrix $P = (p_{ij})$ then defines how likely the underlying is to move from state i to another state j over the next time period. Assuming that these transition probabilities are time-homogeneous, we can write this mathematically as

$$p_{ij} = Pr\left(S_{t+1} = j \mid S_t = i\right) > 0 \quad \forall\, i, j \leq n,\, t > 0. \tag{3.4}$$

If, given sufficient time, it is possible to reach any state from any other starting state, then P is said to be *irreducible* and it must hold that $p_{ij}^t > 0$ for some t.

[7] Interested readers can find further mathematical detail in Ross (2015). Extensions to the original theory are presented in Carr and Yu (2012), Dubynskiy and Goldstein (2013), Martin and Ross (2013), Walden (2014) and Borovička et al. (2016).

In this work, we will let P represent the transition probability matrix (TPM) defined under the risk-neutral measure. In contrast to the RND, transition probabilities are not directly quantifiable from option prices but rather need to be estimated from a given RND surface. In a similar vein, we will denote the real-world transition matrix as $F = (f_{ij})$, and we define the ratio of risk-neutral to real-world transition probabilities as

$$\psi_{ij} = \frac{p_{ij}}{f_{ij}}. \tag{3.5}$$

This ratio is referred to as the pricing kernel in economics literature (Ross, 1976), the stochastic discount factor in financial economics literature (Cochrane, 2001), and the Radon-Nikodym derivative in option pricing literature (Shreve, 2004). Regardless of its name, $\psi_{ij} > 0$ represents the factor that transforms risk-neutral transition probabilities into their real-world counterparts.[8] This also mathematically illustrates the point made earlier in Section 2; namely, that a change in risk-neutral probabilities does not automatically imply a change in real-world probabilities.

Equation 3.5 also makes it clear that one needs to solve for two unknowns simultaneously in order to recover the real-world probabilities. In order to do this, we start by assuming that the pricing kernel is transition-independent (i.e. independent of the asset path). This assumption allows us to then define the pricing kernel as

$$\psi_{ij} = \delta \frac{h(S_j)}{h(S_i)}, \tag{3.6}$$

where $h(S_i)$ is a positive function of the states and δ is a positive discount factor. Combining Equations 3.5 and 3.6, we have that

$$p_{ij} = \delta \frac{h(S_j)}{h(S_i)} f_{ij}. \tag{3.7}$$

Recovery of the real-world probabilities thus relies on estimating the values for p, δ and h from the option-implied RND only, which at first glance appears impossible. However, by imposing certain constraints on the matrix P, Ross (2015) shows that this can in fact be achieved. In particular, if one assumes that P is non-negative, irreducible and time-homogeneous, then according to the Perron-Frobenius theorem there exists a unique positive eigenvalue value λ and corresponding unique positive eigenvector $\mathbf{z}$ such that

$$P\mathbf{z} = \lambda \mathbf{z}. \tag{3.8}$$

Letting $H = diag(h(S_1), h(S_2), \ldots, h(S_n))$, we can rewrite Equation 3.7 in matrix notation,

$$P = \delta H^{-1} F H \iff F = \frac{1}{\delta} H P H^{-1}. \tag{3.9}$$

[8] The pricing kernel must be positive to ensure no arbitrage.

Using this expression for F and coupling it with the fact that each row of the real-world TPM must sum to one, we can write

$$\mathbf{1} = F\mathbf{1} = \left(\frac{1}{\delta} H P H^{-1}\right) \mathbf{1}, \tag{3.10}$$

where $\mathbf{1}$ is an n-vector of ones. Finally, we rearrange Equation 3.10 to obtain

$$P\left(H^{-1}\mathbf{1}\right) = \delta\left(H^{-1}\mathbf{1}\right). \tag{3.11}$$

Written in this form, it becomes clear that Equations 3.11 and 3.8 are equivalent if and only if $\mathbf{z} = H^{-1}\mathbf{1}$ and $\lambda = \delta$. What this means practically is that one can obtain all three unknown variables in Equation 3.7 directly from the option-implied P matrix using an eigenvalue decomposition and thus successfully recover the real-world density. Ross (2015) formalises this result in his recovery theorem:

> *If the market is arbitrage-free, if the pricing matrix is irreducible and if it is generated by a transition-independent kernel, then there exists a unique (positive) solution to the problem of finding the natural probability transition matrix, F, the discount factor, δ, and the pricing kernel, ψ.*

3.3.2 Implementing the Recovery Theorem

To date, there have only been a handful of empirical studies on the recovery theorem (Spears, 2013; Audrino et al., 2015; Backwell, 2015; Kiriu & Hibiki, 2015; Tran & Xia, 2015). A common thread running through these studies is that it is very difficult to implement this theorem. The reason for this is because successful recovery requires one to solve two ill-posed problems. The first of these – estimating the RND surface – has been discussed at length in Section 2. The second problem is the estimation of the risk-neutral TPM from the obtained RND surface. In contrast to the RND literature, to our knowledge only the five papers noted above have considered this secondary problem in any level of detail. Given that estimation of the TPM plays such a crucial role in the practical implementation of the recovery theorem, we spend some time below discussing the various aspects of the estimation procedure.

The initial estimation method put forward by Ross (2015) makes use of the assumption of a time-homogeneous TPM to set up a system of linear equations

$$Q'_{1:n,\tau} P = Q'_{1:n,\tau+1}, \tag{3.12}$$

where $Q'_{1:n,\tau}$ denotes the discretised RND of term τ across the specified n states in P. Equation 3.12 means that the RND of term $\tau + 1$ is equivalent to the product of the RND at term τ and the constant TPM. Tran and Xia (2015) show that the state

discretisation specified for the P and Q matrices can materially alter the recovered probability values in certain settings, meaning that setting the state space should be viewed as an active decision in the recovery process. However, they also show that recovery can be consistent across differing state specifications provided that the varying P matrices are consistent in terms of the sum of smaller discretised states adding up to the equivalent larger discretised states.[9]

Letting $A = Q'_{1:n,1:T-1}$ and $B = Q'_{1:n,2:T}$, we can write Equation 3.12 in a standard ordinary least squares (OLS) form,

$$P = \underset{p_{ij} \geq 0}{\operatorname{argmin}} \|AP - B\|_2^2 \,. \tag{3.13}$$

The A and B matrices can be quite large in practice, making direct optimisation of this objective function an onerous exercise. Thankfully, one can recast the problem as a series of independent vector OLS problems which can be solved much more easily and quickly by standard optimisation packages.

In theory then, it would seem that the second ill-posed estimation problem has a fairly simple solution. However, when Spears (2013) attempted to replicate the results originally presented by Ross (2015) using the estimation method given above, the replication was considerably different to the original. This suggests that Ross includes additional constraints on the structure of the transition matrix. To this end, Spears (2013) tested nine alternative constrained estimation methods and, using a range of fitting criteria, found that one needs to impose considerable structure on the transition matrix in order to obtain a solution which is both economically suitable and statistically robust.

Rather than impose constraints directly on the transition probabilities, Audrino et al. (2015) consider the alternative route of using Tikhonov regularisation on the constrained OLS problem (Tikhonov & Arsenin, 1977). In essence, the idea of regularisation is to introduce an additional term into the objective function which penalises the optimisation from estimating a P matrix that is too far away from a predefined target matrix. Classic Tikhonov regularisation uses the null matrix as the target but one can generalise this to any target matrix depending on the type of structure that one wants to impose.

In this vein, Kiriu and Hibiki (2015) select a target transition matrix $\bar{P} = f(Q)$ constructed from the input Q matrix that ensures that the highest transition probabilities are generally found along the diagonal (see Appendix A.2 for construction details). This means that one is assuming that the underlying is more likely to remain in its

[9] Although not shown here, we test several alternative state space grids in our algorithm below and find little difference in the recovered results.

current state than move to a new state. Kiriu and Hibiki thus attempt to solve the following regularised OLS problem:

$$P = \underset{p_{ij} \geq 0}{\operatorname{argmin}} \|AP - B\|_2^2 + \zeta \left\| P - \bar{P} \right\|_2^2, \tag{3.14}$$

where $\zeta > 0$ is the regularisation parameter that governs the weight given to the additional regularisation norm. Setting $\zeta = 0$ returns the original OLS problem.

Although regularisation techniques are often used to very good effect to stabilise the solution set in ill-posed problems, they do introduce the additional issue of selecting the optimal regularisation value, ζ^*. In order to find this optimal value, one needs to introduce a new function that measures the trade-off between solution smoothness and target distance. Common examples include functions based on Euclidean distance (Backwell, 2015), relative entropy (Audrino et al., 2015) or problem-specific selection functions (Kiriu & Hibiki, 2015). After testing each of the respective methods proposed in the above papers, we choose to adopt the selection function proposed by Kiriu and Hibiki (2015) for its appreciably better robustness. See Appendix A.2 for function details.

Having outlined the necessary theoretical and practical issues, we implement the recovery theorem as follows:

1. Estimate standardised RNDs as per the procedure given in Section 2.3 and construct a discrete Q matrix spanning a $50 - 150\%$ range of the prevailing spot level in 5% intervals.[10]

2. Set the TPM period length as three months, in line with the underlying market expiry structure. The input matrices in the OLS problem are thus defined as $A = Q'_{1:21,1:T-3}$ and $B = Q'_{1:21,4:T}$.

3. Construct Kiriu and Hibiki's (2015) regularisation target matrix $\bar{P}$ and solve the regularised OLS problem in Equation 3.14 using the constrained linear least-squares solver in MATLAB® for a wide range of regularisation values, $\zeta = \{0, 10^{-5}, 10^{-4.8}, \ldots, 10^{2}\}$.

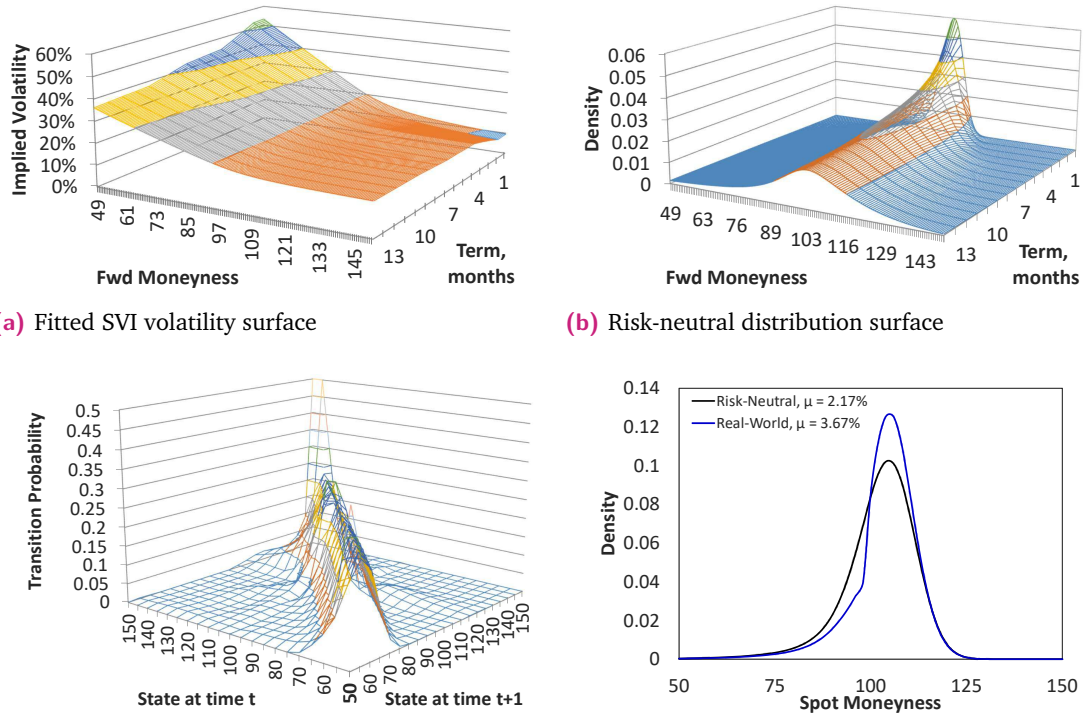
4. Find the regularisation value that minimises the selection function, $\zeta^* = \operatorname{argmin} h(\zeta)$.

[10] We tested increasingly fine discretisation schemes and found consistent recovery results for Q matrices ranging from 21 to 51 states. As a result, we choose the lower bound 21-state estimation to reduce computation time of the regularised OLS problems over the full data sample.

5. Using the optimal ζ^*, solve another regularised OLS problem on a finer 51-state Q-matrix in order to estimate a final risk-neutral transition probability matrix P^*.

6. Use the estimated P^* matrix and the recovery theorem to obtain the three-month real-world transition probability matrix F by applying the Perron-Frobenius theorem, and thereafter extract the discrete three-month real-world return distribution as the middle row of the F matrix.

Figure 3.1 depicts several key outputs from the complete recovery procedure, implemented as at 23 April 2007. Notice the difference in three-month mean estimates under the implied risk-neutral and real-world distributions.



(a) Fitted SVI volatility surface

(b) Risk-neutral distribution surface

(c) Transition probability matrix

(d) Three-month implied distributions

Figure 3.1. Representative portrayal of the real-world probability recovery process for Top40 options data, implemented as at 23 April 2007.

3.4 Top40 Option-Implied Distributions and Tactical Strategies

We use the estimation algorithm outlined in Section 2.3 to create weekly arbitrage-free implied volatility and implied RND surfaces for the Top40 index over the period 5 September 2005 to 16 May 2016, giving a total of 559 surface observations. We then estimate three-month implied real-world distributions using the recovery algorithm given in Section 3.2. Similarly to Audrino et al. (2015), we consider the evolution and correlation of the first four implied moments – mean, volatility, skewness and kurtosis – relative to the underlying asset over the test period and use these moments as input signals in an index/cash timing strategy.

Using the methodology outlined in Sections 3 and 4 above, we estimate risk-neutral distributions across all available terms as well as 3-month real-world distributions for the Top40 index and the USDZAR exchange rate. From these implied distributions, we focus on the evolution and correlation of the first four implied moments – expected return, volatility, skewness and kurtosis relative to the underlying asset over the test period. Finally, and similarly to Audrino et al. (2015), we consider the difference in performance of timing strategies based on recovered real-world measures and implied risk-neutral measures.

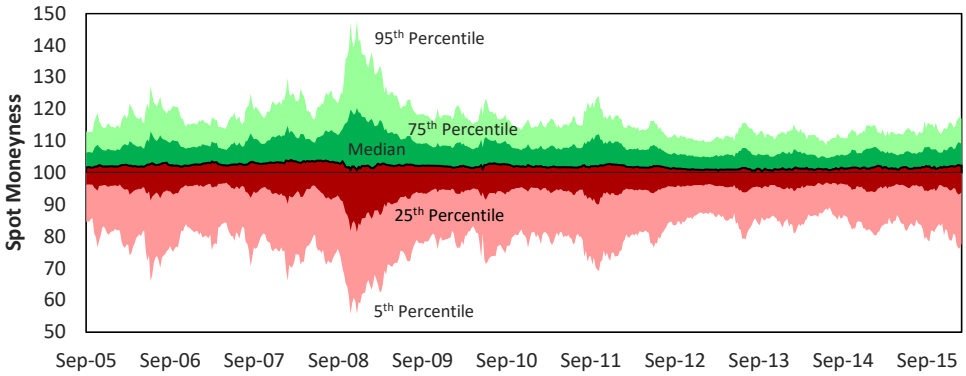


Figure 3.2. Weekly box plot of three-month risk-neutral Top40 distributions, Sep 2005 to Jan 2016.

Before analysing these moments though, we consider the general evolution of the Top40 implied distribution. Figure 3.2 displays a dynamic box plot of the three-month implied distribution over the ten-year sample. This information can be used descriptively or prescriptively. In the descriptive sense, the negative tail of the distribution (red fill) is consistently much longer than the positive tail (green fill) throughout the period, indicative of the larger negative jump or crash risk common to equity indices. The ratio of these two areas thus describes the level of implied

asymmetry in the index at any point in time. The RND widened considerably during the global financial crisis and remained wider than usual until late into 2009. Since then, the RND has narrowed considerably, although the negative tail has once again started to increase over the last couple of years. In the prescriptive sense, one could, for example, focus on the 95th distribution percentile. This is essentially the three-month implied value-at-risk of the Top40 index and can thus be used in a number of forward-looking risk management applications.

3.4.1 Risk-Neutral and Real-World Implied Moments

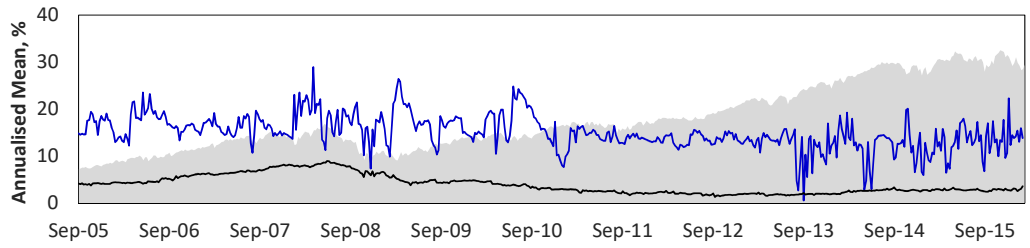
We now consider the implied moments of the risk-neutral and recovered three-month distributions. Figure 3.3 compares the first four risk-neutral moments to their real-world counterparts, against the performance backdrop of the underlying Top40 total return index. Table 3.1 gives the corresponding summary statistics. Notice that the real-world mean is almost always considerably higher than its risk-neutral counterpart – essentially the quarterly cost of carry – and also displays significantly more time variation. Table 3.1 also shows that the annualised real-world mean (and volatility) is very close to the historical annualised Top40 return mean (and volatility) over the sample period. We stress again that recovery of a real-world mean estimate purely from the options market is a remarkable feat and gives one considerable insight into the actual market views used by option market participants for pricing purposes. Coupled with a greater macro view of the derivatives market, this also gives one an inkling of how structural issues such as liquidity and the supply/demand ratio may affect the implied market views used in derivatives pricing.

Table 3.1. Top40 and implied moment summary statistics, Sep 2005 to May 2016

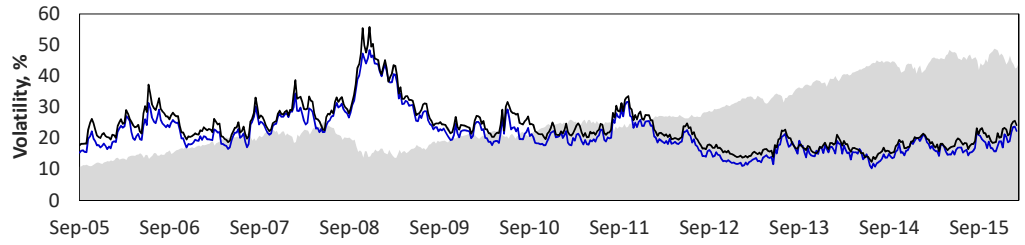
	Mean	Volatility	Min.	5th	25th	50th	75th	95th	Max.
Top40[1]	16.1%	21.5%	-11.7%	-4.7%	-1.2%	0.5%	2.1%	4.6%	15.3%
Risk-Neutral Moments									
Mean[2]	4.0%	1.9%	1.3%	1.8%	2.4%	3.3%	4.9%	7.9%	9.0%
Volatility[2]	23.5%	6.8%	12.4%	15.2%	18.8%	22.1%	26.6%	34.7%	55.8%
Skewness	-0.91	0.29	-1.67	-1.40	-1.12	-0.92	-0.74	-0.45	0.09
Kurtosis	4.95	1.24	2.68	3.20	4.16	4.75	5.62	7.21	10.97
Real-World Moments									
Mean[2]	15.1%	3.6%	0.6%	9.1%	13.1%	14.7%	17.4%	21.2%	28.9%
Volatility[2]	21.3%	6.6%	10.3%	13.1%	17.0%	19.9%	24.2%	32.1%	48.3%
Skewness	-1.05	0.28	-1.71	-1.48	-1.25	-1.07	-0.87	-0.62	-0.08
Kurtosis	5.39	1.34	2.21	3.31	4.53	5.38	6.26	7.67	11.00

[1]Top40 Returns mean and volatility are annualised; percentiles are weekly numbers.

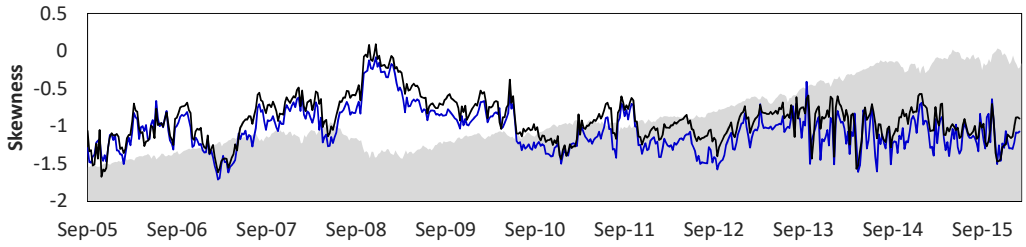
[2]Mean and Volatility moment values are annualised.



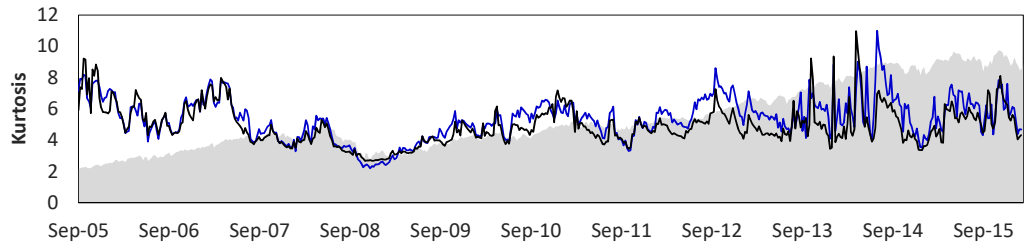
(a) Annualised Mean, %



(b) Annualised Volatility, %



(c) Three-month Skewness



(d) Three-month Kurtosis

Figure 3.3. Top40 real-world (blue) vs. risk-neutral moments (black), with index total return performance as shaded backdrop, Sep 2005 to May 2016.

From the second panel in Figure 3.3, we clearly observe the well-known inverse relationship between asset performance and implied volatility. More importantly though, the two implied volatility profiles are very similar, in line with what one would expect given the theory underlying the BSM pricing framework after accounting for a potential volatility premium added by market makers. This suggests that our recovery algorithm is giving us results that at least match the minimum arbitrage pricing criteria.

The recovered skewness is almost always lower than the risk-neutral skewness, although both are considerably negative as one would expect for three-month index returns. Furthermore, and in line with the literature, both implied distributions are essentially symmetric during the global financial crisis (zero skewness) and only show significant negative skewness again post the market recovery. A similar but inverted relationship is seen between implied kurtosis and index performance during this period. However, there are also several periods in the full ten-year sample when kurtosis declines but the index displays positive performance, making it difficult to generalise this finding without further analysis.

Table 3.2 displays the correlation matrix of changes in the weekly risk-neutral and real-world implied moments. The table has been split into quadrants, which, going anticlockwise, denote correlations between risk-neutral moments only, between risk-neutral and real-world moments, and between real-world moments only. Comparing the upper left and lower right quadrants, one observes that the real-world higher moments have considerably stronger relationships with the lower moments than their respective risk-neutral counterparts. This is particularly noticeable for kurtosis.

Table 3.2. Correlation matrix of weekly implied moment changes, Sep 2005 to May 2016

		Risk-Neutral				**Real-World**			
		Mean	Volatility	Skew	Kurtosis	Mean	Volatility	Skew	Kurtosis
Risk-Neutral	Mean	1.00							
	Volatility	-0.32	1.00						
	Skew	-0.15	0.24	1.00					
	Kurtosis	0.09	-0.13	-0.81	1.00				
Real-World	Mean	0.06	-0.02	-0.12	-0.08	1.00			
	Volatility	-0.23	0.88	0.30	-0.14	-0.33	1.00		
	Skew	-0.13	0.29	0.73	-0.40	-0.35	0.39	1.00	
	Kurtosis	0.09	-0.30	-0.68	0.59	0.46	-0.51	-0.79	1.00

The lower left quadrant displays the significant positive relationship between risk-neutral and real-world volatility, as well as the positive correlations between the two skewness and kurtosis measures respectively. However, there is still evidence to suggest that the informational content available from each pair of moments is different. This is particularly evident when observing the large differences in correlations between the real-world mean and the risk-neutral moments versus the comparative correlations to the other real-world moments.

3.4.2 Tactical Asset Allocation with Option-Implied Information

We heuristically test the forward-looking information content of the moments by following a simple tactical asset allocation (TAA) strategy advocated by Audrino et al. (2015) over the sample period. For expected return, skewness and kurtosis, if the current week's values are greater than the prior week's, then we hold the Top40 index, otherwise we move into cash.[11] We take the opposite strategy for volatility given the well-known inverse relationship with underlying returns. Although simple, this strategy is in line with an investor wanting higher returns, higher skewness and higher kurtosis.[12]

Figure 3.4 displays the cumulative log returns of the strategies versus the Top40 total return in black. The blue shaded lines denote the strategies based on risk-neutral moments, while the red shaded lines denote the strategies based on recovered real-world moments. Table 3.3 gives the summary statistics for all the trading strategies and the Top40 index.

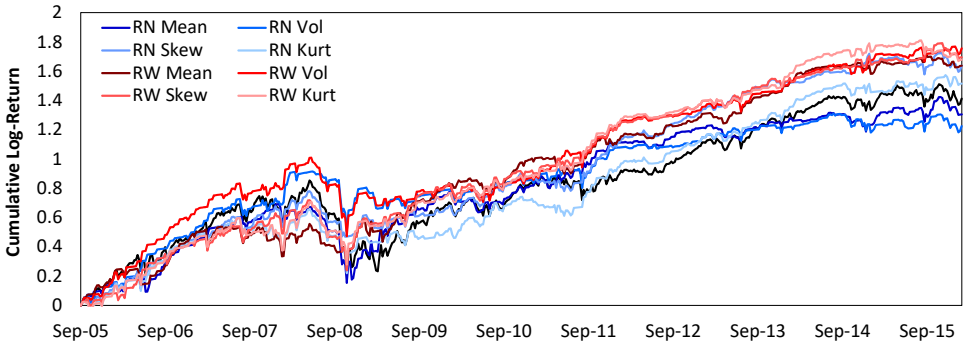


Figure 3.4. Cumulative log performance of implied moment trading strategies versus Top40 index (black line), Sep 2005 to May 2016.

The most striking observation from the depicted and tabulated results is that the recovered moment strategies consistently and considerably outperform the Top40 total return index and the risk-neutral moment strategies, with the exception of the risk-neutral skewness strategy. The average return for the real-world strategies ranges between 16.9% and 18.1%, which is 0.8% to 1.9% higher than the Top40 return over the same period. The volatility of the real-world strategies is also considerably lower than that of the Top40, meaning that the risk-adjusted performance

[11] Transaction costs are not included as we are only interested in assessing the informational content for now. Furthermore, the number of trades is fairly consistent across all timed strategies, meaning that costs would have a similar impact throughout.

[12] Note that an investor will generally only want higher kurtosis and thus fatter tails when skewness is become increasingly positive.

Table 3.3. Top40 implied moment trading strategy results, Sep 2005 to May 2016

	Top40	Risk-Neutral Moment TAA				Real-World Moment TAA			
	Index	Mean	Volatility	Skew	Kurtosis	Mean	Volatility	Skew	Kurtosis
Mean	16.1%	14.3%	13.8%	17.0%	16.7%	16.94%	18.05%	17.51%	17.08%
Volatility	21.7%	17.1%	16.4%	15.3%	16.7%	15.8%	16.0%	15.8%	16.2%
Sharpe	0.50	0.52	0.52	0.77	0.69	0.74	0.80	0.77	0.73
Skew	-0.24	-0.13	0.07	-0.14	0.13	0.28	0.03	0.35	0.23
Kurtosis	5.99	11.37	12.01	8.19	10.41	12.45	12.97	12.17	11.51
Max DD*	-46.2%	-41.6%	-33.6%	-26.8%	-35.2%	-27.0%	-41.6%	-38.0%	-34.8%
# Trades	n.a.	289	243	283	268	296	269	293	291

*Max DD = Maximum Drawdown

of the physical index – as measured by the Sharpe ratio – is consistently lower than the real-world timed strategies. Interestingly, the risk-adjusted comparison between risk-neutral and real-world strategies is not as clear cut. The mean and volatility strategies are clearly dominated by the real-world moments, whereas the comparison is much closer for the higher moment strategies. This suggests two points: firstly, that higher moments are important in a TAA context, and secondly, that the information content within the implied risk-neutral higher moments may be as valuable as that gained from the recovered real-world counterparts. We leave a proper discussion of this conjecture for future research.

Kurtosis of the timed strategy returns is significantly higher than for the index. However, skewness is generally positive as well, meaning that the timed strategies actually display significant positive tail risk. This stems from the fact that one generally moves to cash during some of the worst market downturns, thus decreasing the number and size of negative tail events. This can also be seen in the reduced maximum drawdown numbers relative to the index portfolio.

In summary, the results given above suggest that there is additional information to be gained by recovering real-world implied moments and, furthermore, that this information can be of practical value, at least in the case of tactical asset allocation.

3.5 Conclusion

Given the forward-looking nature of the derivatives market, it is reasonable to surmise that there may be information embedded in option prices. Numerous authors have shown that such option-implied information significantly outperforms the comparative information estimated from price history across a range of portfolio, risk management and trading applications. Although the estimation of risk-neutral option-implied information is well-established in the literature, estimation of the

same in an illiquid market is not. Furthermore, there has been little empirical research done to date – in liquid and illiquid markets alike – on extracting real-world implied information using the recovery theorem introduced by Ross (2015). In this work, we address both of these issues by considering in detail the estimation and application of risk-neutral and real-world option-implied distributions in an illiquid market setting.

We show that the deterministic SVI volatility model is a viable candidate for modelling implied volatility surfaces and use this model to estimate the underlying risk-neutral distributional surfaces on the Top40 index. The issue of calibration with sparse and noisy data is considered at length and a simple but robust fitting algorithm is proposed.

We then describe a robust methodology based on regularised least squares for extracting the implied real-world probabilities and implement this method on a history of weekly SVI implied volatility surfaces for the Top40 index. We discuss how one can use this information descriptively and prescriptively and, furthermore, analyse the recovered moments from the implied distributions. The recovered real-world moments are shown to be in line with economic rationale and also show promising results when used as signals within a simple tactical asset allocation framework.

There exist several avenues for further research based on the work in this chapter. Firstly, it remains an open question as to which of the many proposed techniques is truly optimal for estimating risk-neutral distributions. A large-scale comparative study across a broad range of market settings would thus be very useful for practitioners currently making use of risk-neutral information. Secondly, although there have been a number of theoretical studies on the recovery theorem to date, numerical and empirical research on the topic remains limited. One potential study in this vein could be to examine how adding constraints on the output recovered distribution and pricing kernel affects the empirical recovery process.

4

In Search of the Perfect Hedge Underlying

> *I may not be perfect, but parts of me are excellent.*
>
> **— Ashleigh Brilliant**
> (Author and cartoonist)

Chapter Synopsis

This chapter attempts to answer the practical question, what underlying should one use to hedge an active portfolio? In order to do so, we initially consider three different decompositions for active return and tracking error. These decompositions focus on different aspects of the portfolio, allowing us to precisely quantify the underlying risk and reward drivers in the fund. Thereafter, we describe a general mixed integer programming framework that allows one to select a sub-basket of assets that will most accurately replicate the identified sources of risk and reward whilst simultaneously complying with real-world market constraints. We then study how the effectiveness of an index hedge decreases with a portfolio's tracking error, where effectiveness is measured in terms of the change in downside risk measures of the hedged portfolio.

Motivated by these elements, we introduce several alternative hedging methods for the fund manager to implement a better hedge for their active portfolio. In particular, we consider the use of long-only and long/short basket options as a means of creating more appropriate portfolio hedges. At the same time, we introduce a novel, practical pricing methodology for long/short basket options.

This chapter is adapted from the working paper by Flint et al. (2015a) and addresses research questions 1c, 1d and 4a given in Section 1.2.3.

4.1 Introduction

Derivatives are typically used for two purposes in portfolio management; to implement tactical asset views and thus enhance portfolio returns, or to hedge against particular risk drivers within the portfolio. In this work we focus on the latter application and consider the practical use of derivatives as risk management tools for hedging against systematic market risk. The obvious question is then, what is the optimal hedge for a given portfolio under particular market conditions?

Seymour et al. (2012) suggest that the optimal hedge is that which leads to an overall portfolio with risk and return properties consistent with the investor's preferences, an idea very similar to that proposed in conventional asset allocation. Their research considers optimisation of the hedging structure (e.g. futures, put spreads, collars), the characteristics of the chosen structure (e.g. strike and term), and the proportional size of the hedging structure in the overall portfolio. They also propose a systematic approach for finding such an optimal hedge, again in a manner not dissimilar to how one would determine an optimal asset allocation policy

Another consideration that frequently arises when contemplating hedging is that of timing. In particular, when is it optimal to hedge a portfolio and when should it rather be left unprotected? There will always be a cost attached to option-based portfolio insurance strategies, be it direct or indirect. One therefore needs to be certain that the value of the risk reduction afforded by following such a strategy is higher than its associated costs. Flint et al. (2014), among others, consider the question of when a manager should hedge their portfolio. Motivated by the adaptive market hypothesis introduced by Lo (2004), they outline a systematic process for creating timed hedging strategies based on quantitative market indicators and successfully backtest several dynamic hedging strategies in a South African market context.

This research adds to the existing research by tackling a different aspect of the optimal hedging problem. In particular, this work attempts to answer the question, what is the optimal *underlying* – as opposed to optimal structure – that one should use to hedge an active portfolio? This is a particularly pertinent question for the South African market owing to the recent advent of the Johannesburg Stock Exchange CanDo option platform, which essentially allows one to list, and hence exchange-trade, any exotic option. In particular, listed options on long-only and long/short custom baskets are now readily available to managers as potential portfolio hedging instruments. This has led to an explosion of choice in the hedging candidate underlyings available to the manager. While many of these candidate underlyings will neither be applicable nor perhaps tradable, there will still be a large number that can and should be considered.

Given that the optimal underlying will always be portfolio-specific, this research is not intended to prescribe a particular solution. Rather, the objective here is to describe how one should approach the question of finding the optimal hedge underlying for any given active portfolio. To achieve this goal, we start by considering several methods for decomposing and quantifying the sources of risk and reward within a portfolio. Thereafter, we describe a framework for selecting the subset of assets that best mimic those identified sources. We then tackle the issue of measuring when the standard index hedge is and is not sufficient for an active portfolio. Finally, we introduce several alternative hedging methods for the active manager based on long-only and long/short custom basket options. For the practical components of this research, we focus on application in the South African equity market and consider equity-only portfolios. However, the theoretical concepts and ideas discussed here are applicable to any general active portfolio management setting.

The rest of this chapter is organised as follows. Section 4.2 outlines the fundamental tenets of active portfolio management and provides a brief discussion on the relationship between tracking error and active share. Based on these fundamentals, a framework is developed which allows one to simulate realistic active portfolios via a constrained brute-force algorithm. Section 4.3 introduces several novel active portfolio decompositions that allow one to precisely quantify and therefore manage the risk and reward contributions per active bet. Furthermore, a mixed integer programming approach is presented, which allows one to find the subset of stocks that will most closely replicate a portfolio's future performance while simultaneously complying with real-world market constraints. Section 4.4 then uses the simulated portfolios from Section 4.2 to quantify the effect of mismatch error when using standard index options to hedge increasingly active portfolios. This is done by analysing how the downside risk measures of the hedged returns change with increasing portfolio tracking error. Motivated by these findings, Section 4.5 suggests several alternative hedging methods for active portfolios that provide significantly greater levels of protection. Technical pricing issues are discussed and alternative hedge examples consisting of a long-only and a long/short basket option respectively are given. Section 4.6 concludes and outlines some ideas for further research.

4.2 Active Management Fundamentals

At its core, active management is about making decisions: when to buy or sell any given asset and in what quantity. These decisions are made in order to add value to a passive benchmark, be it a nominated index or cash-based rate. In this setting, value is usually defined in two ways. The first is by achieving a positive return, or alpha, over and above the nominated benchmark at an acceptable level of risk. The

second is by achieving a specified target return at a lower level of risk than that of comparable passive market products.

In both cases, the quality of any active decision taken should be measured by how much value it generates for the fund, conditional on the market and fund constraints faced by the manager at the time. In order to do this rigorously, practitioners generally adopt the framework first articulated by Grinold (1989) and subsequently generalised by Clarke et al. (2002) and De Silva et al. (2006): the fundamental law of active management, or FLOAM. This framework provides the quantitative links between the interconnected areas of signal generation, portfolio construction and underlying market conditions. In doing so, it describes a holistic approach for active portfolio management, both descriptive and prescriptive. We make use of the FLOAM framework in this research in order to generate realistic active equity portfolios, decompose active return and risk, as well as find suitable hedging portfolios.

Let us start by introducing some general concepts and notation which we will use throughout the chapter. Assume that there are N stocks in the underlying investment universe and that the excess-to-cash return R_i on any stock in a given period t is governed by a simple one-factor model:

$$R_{it} = \beta_i R_{mt} + r_t, \tag{4.1}$$

where R_m is the excess-to-cash return on the market, β_i is the sensitivity to the market return and r_i is the independent residual stock return. The first term above represents the systematic component of the stock return and the second represents the idiosyncratic component. The alpha of stock i at time t is then defined as

$$\alpha_{it} = \mathbb{E}\left[r_{it}\right]. \tag{4.2}$$

Dropping the time subscript for simplicity, portfolio returns are then calculated as the weighted sum of underlying stock returns, $R_p = \sum w_{pi} R_i$, and similarly for a nominated benchmark, $R_b = \sum w_{bi} R_i$.

Active portfolio management is generally concerned with active rather than absolute return and risk. Using the framework above, we can define the relationship between *relative* returns, $\Delta R \equiv R_p - R_b$, and *active* returns, R_a as

$$\Delta R = \left(\beta_p - \beta_b\right) R_m + R_a, \tag{4.3}$$

where $\beta_p = \sum w_{pi}\beta_i$ and $\beta_b = \sum w_{bi}\beta_i$ are the market betas of the portfolio and benchmark respectively. We can then also define the relationship between relative risk or tracking error, TE, and active risk, σ_a, in a similar manner:

$$TE^2 = (\beta_p - \beta_b)\,\sigma_m^2 + \sigma_a^2, \tag{4.4}$$

where σ_m is the market volatility. If one considers the special case where the benchmark is the market and the portfolio beta is equal to one, then relative returns are equivalent to active returns and tracking error is equivalent to active risk. This is a common simplification used by practitioners and one which we will use throughout the remainder of the chapter.

Relative returns can also be defined in terms of active weights, w_a, which are the differences in weights between the portfolio and its benchmark:

$$\Delta R = \sum_{i-1}^{N} (w_{pi} - w_{bi})\, R_i = \sum_{i=1}^{N} w_{ai} R_i. \tag{4.5}$$

Assuming that both the portfolio and benchmark weights sum to one, the sum of the active weights must be zero by construction. Therefore, one can separate the active weight portfolio – and thus also the active portfolio return – into an active long portfolio and an active short portfolio of equal weight. Figure 4.1 illustrates an example weight decomposition by graphing benchmark, portfolio, active long and active short weights for a 28-stock portfolio and its 42-stock benchmark.

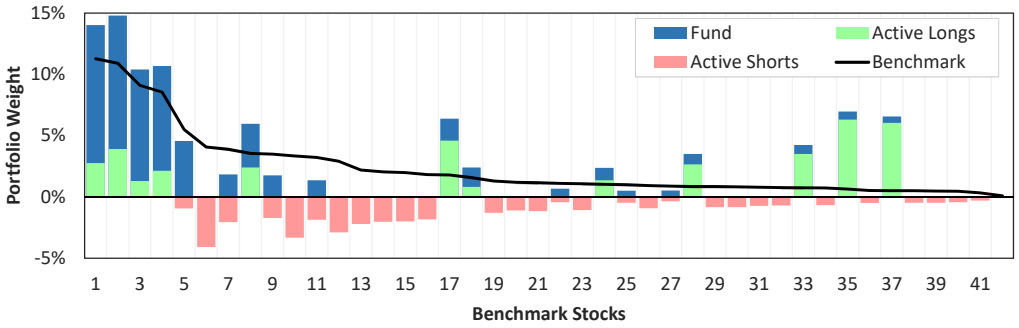


Figure 4.1. Active long and short portfolios of an equity portfolio relative to its benchmark index.

Building from Equation 4.5, we can also redefine tracking error as

$$TE^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ai} w_{aj} \sigma_{ij}, \tag{4.6}$$

where σ_{ij} is the covariance of the returns of stock i and stock j.

In all of the equations above, one can either calculate realised, ex post values from historical data or expected, ex ante values by using forward-looking risk and return expectations. An important distinction pointed out by Hwang and Satchell (2001) between these two calculations is that portfolio and benchmark weights vary

implicitly during ex post calculations, whereas they are taken as fixed during ex ante estimation. Ex post estimation therefore treats portfolio weights as random variables and ex ante estimation treats those same weights as constants. Because of this difference, ex post and ex ante estimates are actually not directly comparable. One practical consequence of this difference in weight treatment is that the ex post tracking error over a given period will always be greater than the ex ante tracking error estimated at the start of the same period. Hwang and Satchell (2001) prove this analytically by deriving the general relationship between ex post and ex ante tracking error. In the absence of covariance estimation error, they show that ex post tracking error can be decomposed into three terms; namely, ex ante tracking error, weighted portfolio weight variance, and the covariance between portfolio weights and asset returns. Hwang and Satchell (2001) show that all three of these terms are always non-negative, thus establishing the result that (fixed weight) ex ante tracking error must underestimate (floating weight) ex post tracking error.

4.2.1 An Active Share Interlude

Another measure of benchmark deviation based purely on active weights which has garnered considerable interest of late is Active Share (Cremers & Petajisto, 2009),

$$AS = \frac{1}{2}\sum_{i=1}^{N} |w_{ai}| \,. \tag{4.7}$$

Active Share is bounded between 0 and 1 and is equal to the size of the active long and active short portfolios respectively. A value of 0 represents an index tracking portfolio, while a value of 1 implies that the portfolio only holds non-benchmark stocks. The interest in Active Share is driven by two features. Firstly, it is considerably simpler to calculate than tracking error as no covariance estimation is required. Secondly, Cremers and Petajisto (2009) assert that Active Share predicts fund performance by showing that the highest Active Share funds in their tested sample significantly and consistently outperformed their respective benchmarks, and vice versa for the lowest Active Share funds. This combination of simplicity and prediction is indeed compelling.

However, several practitioners and academics have recently questioned the validity of using Active Share in a predictive sense (Schlanger et al., 2012; Cohen et al., 2014), arguing that previously reported results suggesting such a conclusion may simply have been driven by the strong correlation between Active Share and the benchmark type (Frazzini et al., 2016). These authors suggest that while positive Active Share is obviously a necessary condition to perform *differently* to a benchmark, it is by no means a sufficient condition for achieving *outperformance*. In fact, Frazzini et al. (2016) contend that Active Share is as likely to correlate positively with performance

as it is to correlate negatively. Thus, while Active Share is clearly an appropriate measure for the size of the active bets taken by a manager, it may not provide any inference as to the skill underlying these active bets. Practitioners should therefore remain wary of conflating Active Share with future manager performance.

Interestingly, and perhaps of more practical use, Sapra and Hunjan (2013) derive an exact relationship between tracking error and Active Share using the FLOAM framework:

$$TE^2 = (\beta_p - \beta_b)^2 \sigma_m^2 + AS^2 \left(\frac{2\pi}{N} \bar{\sigma}_a^2 \right), \tag{4.8}$$

where $\bar{\sigma}_a^2$ is the average residual return variance of all stocks in the universe. Tracking error variance is thus a linear function of squared Active Share, conditional on the given levels of systematic and residual variance. It is clear from Equation 4.8 that if a fund takes large systematic bets relative to stock-specific bets, then Active Share will play a small role in the total tracking error of the fund.[1] However, if one again considers the special case of using the market as the benchmark as well as a portfolio with unit beta, then tracking error variance is directly proportional to squared Active Share. Specifically, the sensitivity of tracking error to Active Share increases for smaller portfolios (smaller N) and during times of greater market dispersion (larger $\bar{\sigma}_a^2$).

4.2.2 Simulating Active Portfolios

In order to properly analyse the effect of mismatch error when using index options to hedge increasingly active portfolios, one first needs to construct a comprehensive range of increasingly active portfolios. However, creating realistic random active portfolios is not a trivial exercise. While a common approach is to follow the generalised alpha generation process of De Silva et al. (2006) based on the market covariance and an assumption of manager skill, this is more suited to finding *optimal* active weights given a specific portfolio objective. Instead of optimal portfolios though, we want to generate the total range of *possible* active portfolios that meet a given set of fund constraints. Therefore, instead of using the alpha signal generation process, we focus directly on active weight generation. In particular, De Silva et al. (2006) show that under the FLOAM framework, the optimal unconstrained active weights are normally distributed with zero mean and variance proportional to the active portfolio risk. Using this as a starting point, we generate realistic random active portfolios based on the following constrained brute-force algorithm:

[1] While we have defined the systematic component above in terms of a single market factor, it is a fairly trivial exercise to show that a similar expression holds when using a multi-factor return model.

1. Specify a portfolio benchmark and stock universe, as well as a covariance matrix for the chosen universe.

2. Specify the target tracking error level, the allowable cardinality range (i.e. number of stocks) and the maximum individual weight for the active portfolio.

3. Randomly select a set number of stocks from the universe and generate active weights from a normal distribution, where the mean fluctuates randomly around zero and the variance is a function of the specified tracking error target level.

4. Scale the generated random weights to match the target tracking error level within a given tolerance, while obeying the maximum weight and portfolio budget constraints.

5. Repeat steps 3 and 4 until a random portfolio is generated which obeys the given constraints.

6. Repeat steps 3 to 5 to generate a large number of portfolios for a given tracking error target.

7. Repeat steps 2 to 6 across a range of tracking error targets.

Table 4.1 details all the variables used in the active weight generation process along with their allowed values or ranges. For the purposes of this study, we limit ourselves to a tracking error range of 0% to 10% and a stock selection universe of either the largest 40 or 100 stocks by market capitalisation on the Johannesburg Stock Exchange (JSE). The FTSE/JSE Top40 (Top40) and FTSE/JSE Shareholder Weighted Top40 (Swix40) indices are specified as the potential benchmarks to ensure that tracking error is always specified relative to the available index hedging instruments in the South African market.

The Current covariance matrix is estimated from three years of daily stock return data ending in August 2015, while Turbulent and Quiet covariance estimates are calculated from daily return data from July 1995 to August 2015 that has been partitioned into two regimes based on the financial turbulence index proposed by Chow et al. (1999).[2] Stock return histories of varying lengths are dealt with by using an initial pairwise correlation and volatility calculation linked with a subsequent

[2]The turbulence index is calculated using total return data from ten South African sector indices and is equivalent to a squared Mahalanobis (1936) distance. Smoothed daily turbulence scores are calculated and used to classify the market as either Turbulent or Quiet based on a 75^{th} percentile cutoff value. See Sections 6.2.4 and 7.2.1 for more detail on financial turbulence.

Table 4.1. Active portfolio simulation variables and specified ranges

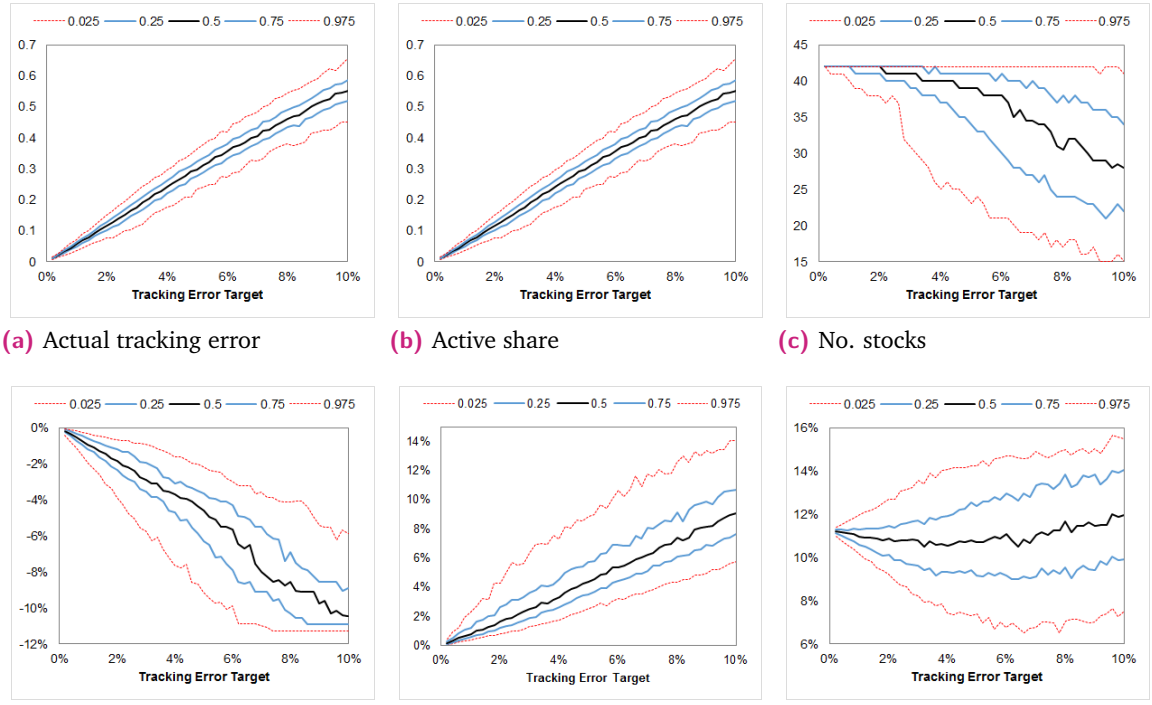
Variables	Values/Ranges
Active Weight	$w_a = w_p - w_b \sim \mathcal{N}\left(\mu_a, \sigma_a^2\right)$
	$\mu_a \in [-2.5\%, 2.5\%]$
	$\sigma_a = \begin{cases} 1.5\% & \forall TE < 3\% \\ \frac{TE}{2} & otherwise \end{cases}$
Benchmark	$Bmk = \{\text{Top40}, \text{Swix40}\}$
Tracking Error Target	$TE \in [0.2\%, 0.4\%, \ldots, 10\%]$
Tracking Error Tolerance	$\epsilon = \lvert 0.025 \times TE \rvert$
Maximum Weight	$w_p \leq 15\%$
Cardinality Range	$K_{T40} \in [15, 42]$
	$K_{T100} \in [25, 100]$
Covariance Estimate	$\Sigma = \begin{cases} \Sigma_{current} \\ \Sigma_{quiet} \\ \Sigma_{turbulent} \end{cases}$
No. Portfolios per TE	$K = 500$

correlation eigenvalue filtering function in order to create valid covariance estimates for each regime. As it turns out, the Current and Quiet regimes are very similar in terms of underlying market conditions. Therefore, only the Current and Turbulent covariance matrices are considered in our tests.

Note that the cardinality range and maximum weight limits given in Table 4.1 are 'soft' constraints in that they may be violated due to the nature of the portfolio generation algorithm. That being said, the number and size of these violations is generally quite small and thus of little practical concern.

Figure 4.2 gives the distributional output from a simulation of 500 active portfolios per TE level from a universe of Top40 stocks using Turbulent market conditions.

In order to achieve a low tracking error of 0.2% - the left-most point in each panel – one needs to essentially hold the benchmark. This is clearly evident in all panels of Figure 4.2 as all the distributions tend towards single values. As one increases the tracking error target, the range of possible active portfolios increases in kind. Given our use of the simplifying market benchmark and zero active beta assumptions, one observes that median Active Share increases quadratically with tracking error, in line with the relationship shown by Sapra and Hunjan (2013). Dispersion in the cardinality and weight distributions respectively also largely increases with tracking error although skewness depends on the statistic under review. The maximum number of constituents and the maximum active short weight are naturally bounded by the choice of benchmark and universe respectively. This results in the hard limits clearly visible in Panels 4.2c and 4.2d. In contrast, we note the soft maximum weight



(a) Actual tracking error (b) Active share (c) No. stocks

(d) Max active short (e) Max active long (f) Max portfolio weight

Figure 4.2. Simulated active portfolio distribution percentiles across tracking error range, estimated under Turbulent market condition.

bound of 15% evident in the maximum active long and portfolio weight distributions shown respectively in Figures 4.2e and 4.2f.

Based on the results given in Figure 4.2, we believe that a sample of 500 random portfolios per tracking error target value – 25 000 portfolios in total – provides one with sufficient coverage of the true distribution of possible realistic active portfolios. For the analysis presented in Section 4.4, we thus make use of the portfolios displayed above as well as another set of 25 000 random portfolios generated using the same process and parameter values but with tracking errors calculated using the Current covariance matrix.

4.3 Selecting an Appropriate Hedge Portfolio

As mentioned in Section 4.1, Seymour et al. (2012) outline a systematic approach for determining the optimal equity hedge for a given portfolio in terms of the type of hedging structure, the characteristics of the chosen structure, and the proportional size of the hedging structure in the overall portfolio. While this approach includes flexibility on several important aspects of the chosen hedging instrument, it does

not necessarily account for the fact that active, non-index portfolios will always be imperfectly hedged by available index derivatives.

Seymour et al. (2014) attempt to address this issue and show how fund managers can limit possible negative return contributions from their active positions by rather using single stock derivative overlays. They also introduce the idea of using option structures on a custom basket of stocks to more accurately hedge out active portfolio risk. Using a high TE and a low TE active portfolio, Seymour et al. (2014) show that downside risk is reduced more by using the appropriate basket hedge rather than the approximate index hedge and that this reduction is significantly larger in the case of the high TE portfolio. However, the two active portfolios analysed by Seymour et al. (2014) are simply made up of the largest ten stocks in the Swix40. For these example portfolios, the choice of hedge portfolio is thus straightforward; simply use the portfolio itself as the underlying basket. However, in many cases the choice of hedge portfolio is not straightforward.

Real-world portfolios generally have far more than ten constituents and some of these counters may well be fairly illiquid. As a result, it may be difficult for market makers to write derivatives directly on the full active portfolio because of their inability to accurately delta hedge the portfolio's illiquid constituents. One therefore needs to take into account the portfolio size as well as the portfolio weights when attempting to create a tradable hedge portfolio. In the South African equity market, liquidity can be a particularly concerning issue. As a result, local practitioners generally use a rule of thumb that the maximum nominal per stock should not exceed 25% of the respective average daily volume traded.

In this research, we extend the initial exploratory analysis of Seymour et al. (2014) by considering the generalised problem of selecting the most appropriate hedge portfolio for any given active portfolio under real-world trading constraints. More specifically, our goal is to find a tradable subset of the portfolio universe that tracks the active portfolio sufficiently well, and thus provides one with a more appropriate hedge underlying than the standard index underlying. There are several ways to approach this problem and we consider one such method in Section 4.3.2. Before describing this method though, we consider the related and more general issue of decomposing active risk and active return.

4.3.1 Return and Risk Decompositions

The objective of portfolio decomposition is to understand what return or risk factors a portfolio is exposed to and in what quantities. This allows one to understand exactly how each portfolio component affects the whole and thus pinpoint exactly which

components are most important. Portfolios can be decomposed in many different ways. In Section 4.2, we have already extensively discussed weight decomposition relative to a benchmark, as illustrated in Figure 4.1. We consider here two further decompositions; namely, active return and tracking error decomposition.

Active Return Decomposition

As described in Section 4.1, the return of a portfolio is equal to the sum of the constituent stock returns weighted by their respective portfolio weights. Similarly, the relative return – taken as the active return under our simplifying assumptions – is equal to the sum of the constituent returns weighted by their respective active weights. But as Macqueen (2011) points out, this is actually just one of way of defining active return:

$$\Delta R = \sum_{i=1}^{N} w_{ai} R_i \tag{4.9}$$

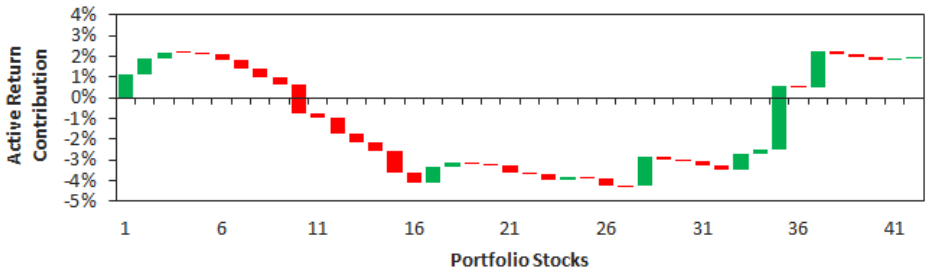
$$= \sum_{i=1}^{N} w_{pi} \left(R_i - R_b \right) = \sum_{i=1}^{N} w_{pi} r_i \tag{4.10}$$

$$= \sum_{i=1}^{N} w_{ai} r_i . \tag{4.11}$$

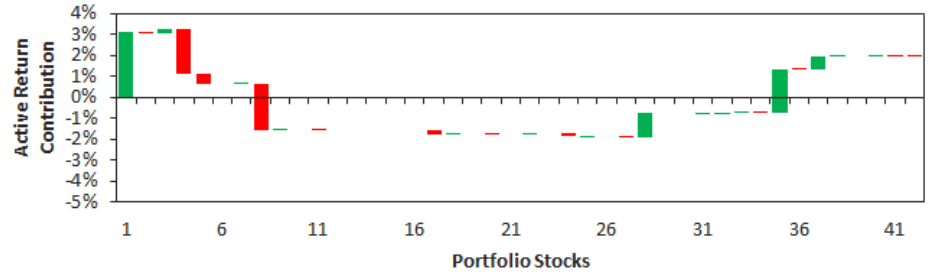
Equations 4.9 to 4.11 formulate active return in three ways: using active weights and total stock returns, using portfolio weights and active stock returns, and using active weights and active stock returns. Each formulation emphasises different aspects of active portfolio return and while the first is probably the industry standard, it is likely not optimal in all situations. As an example, consider the 28-stock active portfolio shown in Figure 4.1. Using expected return estimates from the Current regime, the portfolio has an expected return of 21.71% compared to the benchmark's 19.81%, giving an expected active return of 1.91%. Using the three active return decompositions given above, we calculate stock return contributions and display the different active return contribution paths in Figure 4.3. The contribution path is a means of visually displaying the size and sign of each stock's contribution – as per the size and colour of each bar – but in a cumulative manner so that the ending level of the right-most bar represents the sum of all stock contributions; in this case the portfolio's 1.91% active return.

Clearly, all three formulations give the same portfolio active return as depicted by the matching right-hand end points. However, the individual stock contributions can differ significantly, and this is not just a scaling issue. Stocks with positive contributions in one decomposition can have negative contributions in another.

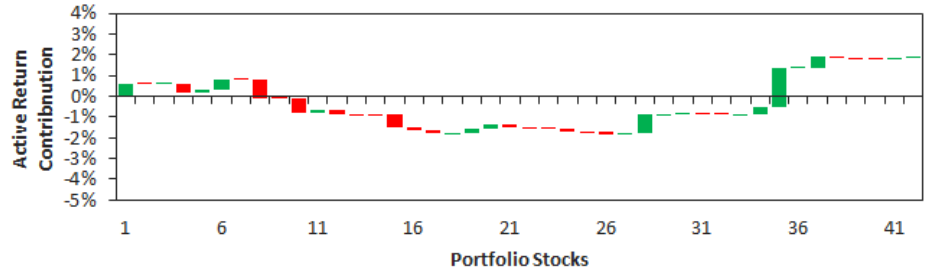
For instance, the return contribution from the sixth stock is negative in the first decomposition, zero in the second and positive in the third.



(a) Active weights and total returns



(b) Portfolio weights and active returns



(c) Active weights and active returns

Figure 4.3. Return contribution paths for different active return decompositions.

Because Figure 4.3a uses full stock returns and active weights, this contribution path highlights how effective all stock selection decisions were relative to the benchmark. Positive contributions in this path therefore come from either overweighting positive performers or underweighting/excluding negative performers. The former can be thought of as an 'explicit' effect, while the latter is an 'implicit' effect. In contrast, Figure 4.3b uses full portfolio weights and active stock returns. This means that all non-portfolio stocks are ignored – and therefore all implicit contributions as well – highlighting only how effective the manager was at taking positive active bets. Finally, Figure 4.3c uses both active weights and active returns. Because of this, the total range of the contribution path is smaller than those given in Figures 4.3a and 4.3b. Positive contributions are now harder to attain because both active weight and active return need to be positive. Macqueen (2011) contends that this decomposition

is the best and most intuitive because it combines the explicit and implicit active weight effects with the effectiveness of the manager at selecting only positive active bets. This path therefore highlights the true contributions of each decision, relative to the benchmark weights as well as the underlying return opportunity set.

Tracking Error Decomposition

From the active return formulations given above, one can also create different tracking error decompositions:

$$TE^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ai} w_{aj} \sigma_{ij} \tag{4.12}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} w_{pi} w_{pj} \tilde{\sigma}_{ij} \tag{4.13}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ai} w_{aj} \tilde{\sigma}_{ij}. \tag{4.14}$$

where $\tilde{\sigma}_{ij}$ is the covariance of the active returns of asset i and asset j. This active covariance matrix can be calculated in a straightforward manner from the full return covariance matrix, the given benchmark weights and the observation that a stock's active return, r_i, can be written as a weighted linear sum of the full stock returns:

$$r_i = R_i - R_b = (1 - w_{bi})\, R_i - \sum_{j \neq i}^{N} w_{bj} R_j. \tag{4.15}$$

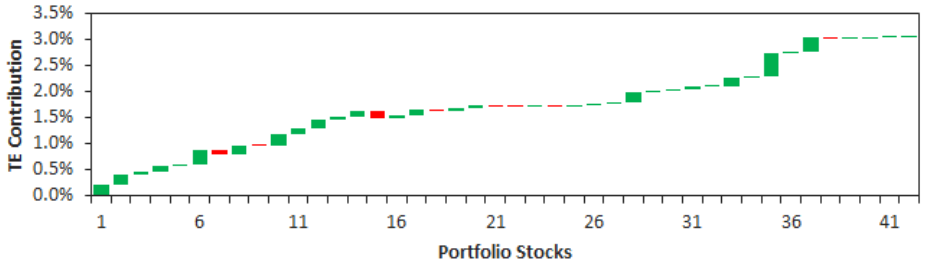
Because covariance is a bilinear function, the active covariance between stocks x and y can similarly be defined as a function of the full covariance matrix

$$\tilde{\sigma}_{xy} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{xi} w_{yj} \sigma_{ij}, \tag{4.16}$$

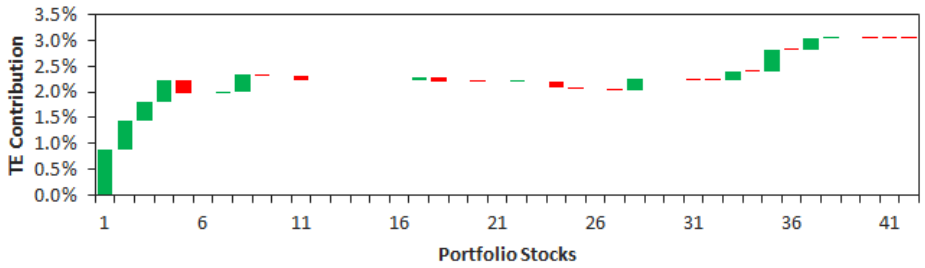
where w_{xi} represents the active weights of a portfolio that only contains stock x. Interestingly, it is evident from Equation 4.16 that a change to any value within the full covariance matrix will cause all active covariance values to change, and this change will be proportional to the pairwise product of the benchmark weights.

As with the active return decompositions, the three tracking error decompositions given above will always give equivalent portfolio tracking error values but can differ on the individual stock level contributions. Consider again the 28-stock active portfolio from Figure 4.1, which has a tracking error of 3.05% under the Current regime covariance. Figure 4.4 shows the respective tracking error contribution paths

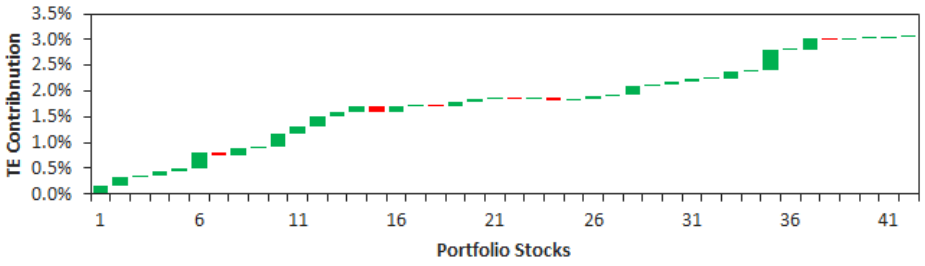
under the three different formulations.[3] Keep in mind that equal contributions from all stocks would be graphed as a straight line ending at 3.05%.



(a) Active weights and total returns



(b) Portfolio weights and active returns



(c) Active weights and active returns

Figure 4.4. Tracking error contribution paths for different tracking error decompositions.

Similarly to active return, each tracking error decomposition emphasises different aspects of a manager's active risk bets, making each contribution path useful in its own right. Figure 4.4b – using portfolio weights and active covariance – ignores non-portfolio stocks and is significantly different from its counterparts. In contrast, Figures 4.4a and 4.4c look very similar to each other, showing only slight differences in some contribution values. These two decompositions should react differently to changes in underlying volatilities and correlations though. Both decompositions use active weights, meaning that explicit and implicit bets are included in the contributions. However, because Figure 4.4c uses the active covariance matrix, the effect of a single volatility or correlation change feeds through to all active covariance estimates, generally leading to more stable tracking error contributions.

[3] Meucci (2009) develops a similar tracking error concentration curve as those shown in Figure 4.4 but calculated using principal portfolios rather than stock positions.

In summary, the active return and tracking error decompositions displayed here allow one to precisely quantify and therefore manage the risk and reward contributions per active bet. This can either be done by repositioning the stock portfolio to target a specific profile or by using derivatives to remove undesirable risk and enhance desirable upside potential.

4.3.2 A Mixed Integer Programming Solution for Selecting the Hedge Portfolio

Let us reiterate the problem posed in the introduction of Section 4.3; how to select an appropriate hedge portfolio for a given active portfolio. While this problem has not been officially addressed in the derivative hedging literature, it has been extensively studied from an index tracking perspective. In this formulation, Beasley (2013) defines the goal of index tracking as identifying the best subset of stocks to hold, as well as their appropriate weightings, in order to replicate the future performance of that index over a given investment horizon. This is nearly exactly the same goal as trying to select the optimal hedging portfolio, with the exception that what is being replicated now is an active portfolio rather than an index. Given this similarity, it would seem obvious to consider the approaches used in the index tracking space.

One such approach which has gained popularity with the advent of increased computing power is mixed integer programming. A mixed integer program is one in which some variables are continuous while others take on integer values. This is ideal for setting up a problem in which one chooses a small subset of stocks from a larger universe – the integer variables – and then searches for the set of weights – the continuous variables – that minimises an objective function under given a set of constraints. The choice of objective function defines the problem either as linear or nonlinear. In general mixed integer programs can be quite hard to solve unless one can formulate the problem in a very particular way. Thankfully, one can do just this for index tracking problems. Below, we discuss a mixed integer linear programming (MILP) and a mixed integer quadratic programming (MIQP) approach for selecting the hedge portfolio which can be solved fairly easily – albeit slowly – with freely available optimisation toolboxes and heuristic solvers. All results below are produced using the YALMIP toolbox for MATLAB® (Löfberg, 2004) in conjunction with MOSEK optimisation software.

Canakgoz and Beasley (2009) propose an MILP formulation of the index tracking problem which includes transaction costs, a portfolio cardinality constraint, and a portfolio turnover constraint. They show that it is possible to view index tracking from a regression standpoint. That is, if a perfect tracking portfolio was regressed against the index, then one would expect to find unit beta and zero alpha. Using

this insight, Canakgoz and Beasley (2009) set up a two-stage MILP formulation that initially solves for a portfolio with unit beta to the index and then solves for a portfolio with zero alpha to the index while maintaining the optimised beta found in the first stage. The major advantage of this approach is that the problem remains linear, meaning that it is computationally easy to solve even for indices with thousands of constituents. As an example, Canakgoz and Beasley (2009) show that under realistic transaction cost limits, a portfolio with as few as 70 stocks can be timeously found that replicates the Russell 3000 to a high degree of precision.

An alternative approach to that proposed above is to find a constrained portfolio that minimises some return dispersion measure relative to the index. The most common such measure is the classical tracking error, although many alternative deviation statistics have been proposed in the literature. Cesarone et al. (2014) and Xu et al. (2016) outline such a MIQP formulation and include cardinality and weight range constraints. While this formulation is more difficult to solve than the linear regression MILP problem, due to its special structure it can still be solved for portfolios with several hundreds of variables.

In this work, we showcase the MIQP framework because of its ties to the more commonly used tracking error measure, and formulate the MIQP hedge portfolio problem as follows:

$$
\begin{aligned}
&\underset{x}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{j=1}^{N} \left(x_i - w_{pi}\right)\left(x_j - w_{pj}\right)\sigma_{ij} && && (4.17)\\
&s.t. \quad \sum_{i=1}^{N} x_i = 1 && (budget\, constraint)\\
&\qquad \sum_{i=1}^{N} z_i = K && (cardinality\, constraint)\\
&\qquad l \leq x_i \leq u \;\; \forall i && (weight\, constraint)\\
&\qquad z_i \in [0,1] \;\; \forall i
\end{aligned}
$$

The x_i are the hedge portfolio stock weights, K is the allowed number of stocks, l and u are the lower and upper weight bounds respectively and z_i is a Boolean vector identifying which stocks are included in the hedge portfolio. This formulation is fairly general and can easily be modified to consider alternative objectives such as minimum variance or equal-weight optimised hedge portfolios.[4]

To illustrate the effectiveness of the MIQP approach, consider the problem of finding an optimal Top40 hedge portfolio under varying hedge portfolio cardinality con-

[4]One can also include a return constraint in the above program if wanted. An example of such an MIQP problem would be enhanced indexation, where the goal is to achieve a fixed positive excess return relative to the benchmark (Beasley, 2013).

straints. No weight range limits are imposed apart from the standard long-only constraint. Figure 4.5 graphs the hedge portfolio's cardinality against its tracking error calculated under Current and Turbulent market conditions respectively, and Figure 4.6 displays the corresponding hedge portfolio weights calculated under Current market conditions.

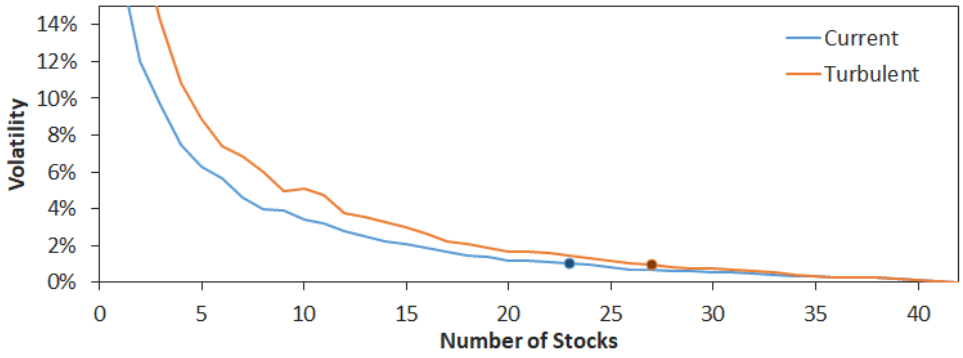


Figure 4.5. Hedge portfolio tracking error versus portfolio cardinality in Current and Turbulent markets.

We define the optimal hedge portfolio as the smallest subset of stocks with a tracking error of 1% at most. Under this definition, we find that one would need 23 stocks to accurately hedge the Top40 under Current conditions and 27 stocks under Turbulent conditions. Out of interest, if one changed the objective to minimise portfolio variance rather than tracking error, one would only require 10 stocks to achieve this goal under both regimes.

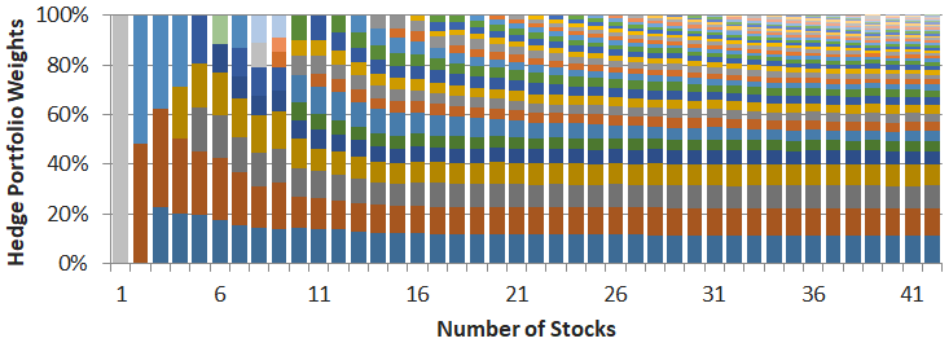


Figure 4.6. Hedge portfolio weights versus portfolio cardinality in Current markets.

Although the example given here may seem somewhat contrived, it is actually a decent test of the effectiveness of the mixed integer programming approach. The Top40 index is arguably more concentrated than most active portfolios would be, making it more difficult to track. Furthermore, testing under Turbulent market conditions – characterised by higher volatilities and correlations – also showcases the robustness of the MIQP approach. The results are generally very stable across the stock range and, although not shown here, have been tested successfully on active portfolios of up to 100 stocks.

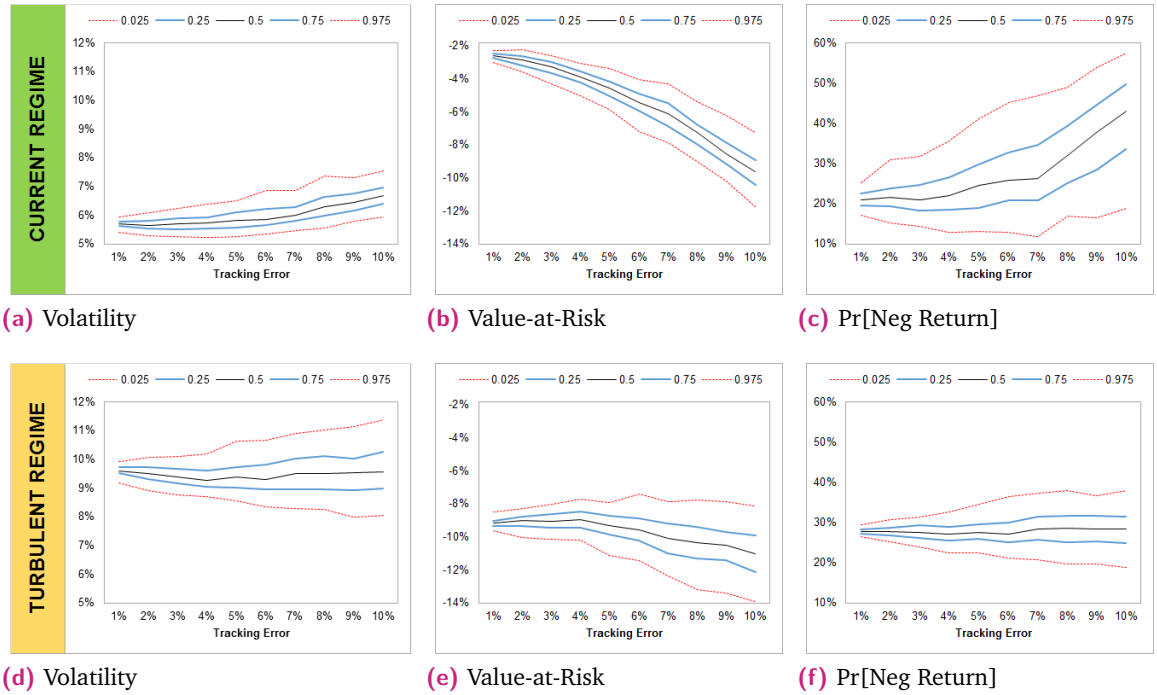
4.4 Hedge Mismatch for Active Portfolios using Index Options

While tracking error provides a well understood measure of divergence between a portfolio and its benchmark index, it is difficult to translate this directly into a measure of hedge mismatch. One approach to achieve this translation is to compare how the hedged return distribution changes as one considers increasingly active portfolios. In particular, by examining the differences in the risk measures of the hedged portfolios, one can quantify the relationship between tracking error and hedge mismatch. This in turn helps one to decide whether an index hedge will provide sufficient protection for a given active portfolio.

Using the portfolio simulation framework outlined in Section 4.2, we analyse the distributional statistics of increasingly active portfolios hedged firstly with a three-month Top40 outright put struck at the prevailing index level and, secondly, with a Top40 put spread with strikes set equal to 100% and 90% of the index level. The three-month expiry hedged return distributions are simulated using both Current and Turbulent market conditions to give an indication of how market regimes affect the severity of the hedge mismatch. We first consider results for active portfolios hedged with an outright put option. Figure 4.7 displays how the volatility, value-at-risk (VaR) – i.e. the 5^{th} return percentile – and negative return probability (NegProb) statistics of the expiry hedge distributions change across the tracking error spectrum.

Although we do see an increase in Current regime median volatility and volatility dispersion across the full tracking error range, this effect is quite small. Median volatility only increases by 1% over the 10% TE range, and the 95% volatility range even at the highest tracking error is still only 1.6%. In comparison, median volatility is much higher under the Turbulent regime but remains constant across tracking error. As expected though, volatility dispersion increases significantly, leading to a maximum 95% range of 3.3%. Note that even for a 10% tracking error, the lowest Turbulent volatility percentile remains above the highest Current regime percentile, emphasising the effect of underlying regime on hedge outcome.

In contrast to volatility, which is a symmetric measure of risk, downside risk measures give one a better indication of the true effectiveness of a given hedge. Looking at the VaR and NegProb panels, it is clear to see the waning protection that an index option provides for active portfolios. In the Current regime, median VaR decreases sharply from -2.8% to -9.8% across the tracking error range, with the worst case VaR – the 'VaR-of-VaR' if you will – being as low as -11.9%. The magnitude of these VaR numbers would likely be too large most managers, especially given the



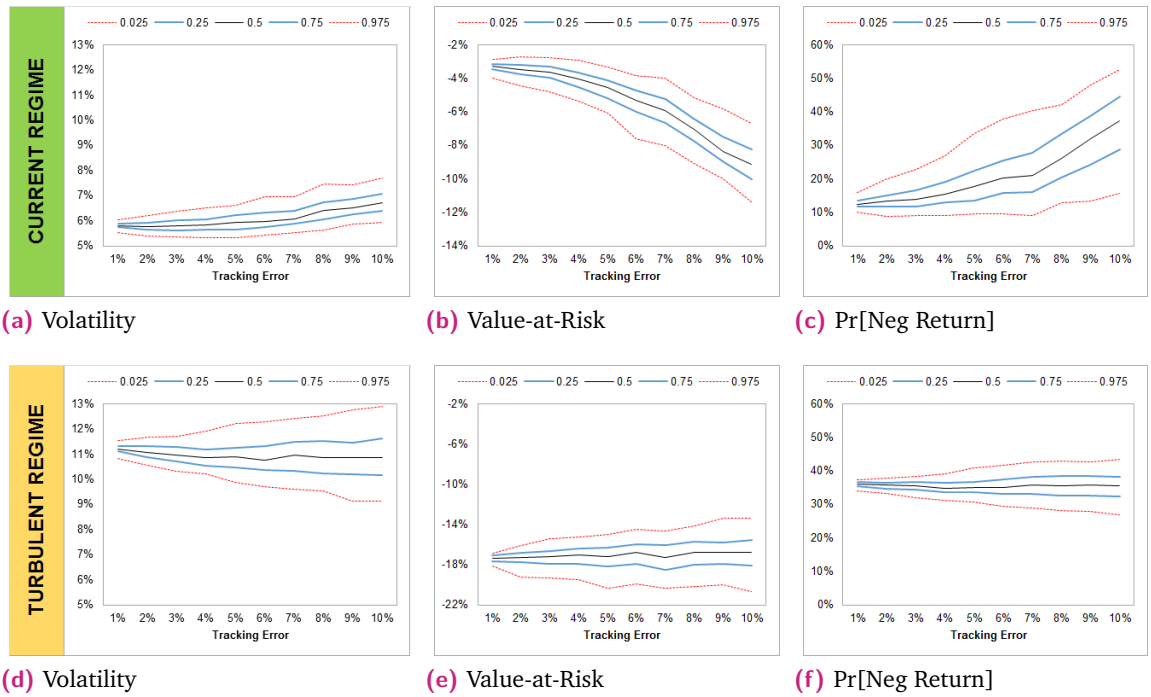
(a) Volatility **(b)** Value-at-Risk **(c)** Pr[Neg Return]

(d) Volatility **(e)** Value-at-Risk **(f)** Pr[Neg Return]

Figure 4.7. Risk distribution percentiles at expiry for 3-month outright put hedged portfolios versus tracking error.

choice of an outright put option as the hedging instrument. A similarly worrying picture is shown for Current NegProb, which starts at a median 21% probability of achieving a negative quarterly return for an index-like portfolio and increases to a 41% probability for very active portfolios. There is also a sharp increase in the right skew of the NegProb distribution. The lower percentile starts at 17%, decreases to a low of 12% at 7% tracking error and then increases thereafter to end at 19%. In comparison, the upper NegProb percentile increases sharply and consistently from 25% to a staggering 57%. This is due to the combination of portfolio composition affecting the portfolio expected return and higher tracking error affecting the hedge efficacy. Finally, note that all NegProb values are increased slightly due to the initial cost of the outright put option.

The Turbulent downside risk measures provide an interesting comparison. When markets are Turbulent – i.e. highly volatile, down-trending and strongly correlated – the median VaR and NegProb values are significantly worse for an index-like portfolio, starting at -9.4% and 28% respectively for 1% tracking error. However, in comparison to the Current risk measures, which show significantly declining hedge efficacy, median Turbulent VaR only falls to -11.2% and median NegProb remains constant at 28%. VaR dispersion does increase more in Turbulent markets, with the VaR-of-VaR now reaching a low of -14.1%. Interestingly though, NegProb dispersion is only half of that seen in Current markets and the distribution remains symmetric across the tracking error range. This is because the general market down-trend



(a) Volatility (b) Value-at-Risk (c) Pr[Neg Return]

(d) Volatility (e) Value-at-Risk (f) Pr[Neg Return]

Figure 4.8. Risk distribution percentiles at expiry for 3-month put spread hedged portfolios versus tracking error.

increases the initial median NegProb, while the increased market volatilities and correlations decrease NegProb dispersion and skewness.

Results for the expiry distributions of the put spread hedged portfolios are given in Figure 4.8. In most panels, the patterns and value ranges are very similar to the outright put risk statistics. However, there are some differences which are worth highlighting. Firstly, given that the put spread only provides protection over the $90\% - 100\%$ index range, the absolute values of the Turbulent risk statistics are considerably higher than those shown for the outright put. Median volatility, VaR and NegProb percentiles remain fixed around 11%, -17% and 35% respectively. Secondly, the put spread VaR distributions are slightly lower for index-like portfolios than their counterparts shown in Figure 4.7 but are slightly higher for very active portfolios. Thirdly, put spread NegProb distributions are also slightly lower – a function of the lower relative cost of the put spread – and show less dispersion than those given in Figure 4.7. This is again a function of the limited protection range of the put spread.

4.5 Alternative Hedging Methods

Section 4.4 showed us that, under certain market conditions, index options may not provide sufficient protection for even moderately active equity portfolios. In

such cases one then needs to consider alternative hedging methods that do provide the required level of protection. At a high level, we can classify the hedge methods available to portfolio managers as follows:

1. Index-only hedge – this is the status quo

2. Index hedge plus single stock hedge overlays

3. Custom long-only basket hedge

4. Index hedge plus active long/short basket hedge

Note that this list is given in decreasing order of hedge mismatch but increasing order of complexity. An index-only hedge is simple to implement but leaves the active component of the portfolio unhedged, as shown above. Assuming that options are available on the underlying stocks, an alternative hedge method would then be to augment the index-only hedge with single stock overlays. In particular, this would mean buying downside protection (i.e. put options) for long active positions and upside participation (i.e. call options) for short active positions. Because these positions are supplementary, one can tailor the type and size of each hedge structure based on the respective conviction levels of the underlying stock bets. That being said, if the goal is to achieve maximum protection for the total portfolio, one is still subject to pricing mismatch between the index and single stock positions as well as potentially severe trading constraints for the single stock overlays.

The third alternative is to directly consider a hedge on a custom basket of stocks. The level of hedge mismatch is thus precisely defined by the manager through the composition of the basket. It is worth keeping in mind though that all custom basket options will also be subject to certain nominal and liquidity constraints, as discussed in Section 4.3.

The fourth alternative is arguably most in line with the FLOAM framework as it is a combination of an index hedge plus a custom basket hedge directly on the active long/short portfolio. As with the custom long-only basket hedge, the level of hedge mismatch can be completely controlled by the portfolio manager. However, it is also the most complex of the hedge methods in terms of pricing and thus would surely carry with it an additional pricing premium.

Obviously, the optimal hedge method will depend on a number of factors, including the level of portfolio activeness, the manager's return and risk objectives, and the prevailing market conditions at the time of hedging. Seymour et al. (2014) discuss several of these issues in the context of single stock overlays. In this work, we

consider specific aspects of the long-only basket hedge method and also provide a general overview of the index plus active long/short basket hedge method.

4.5.1 Basket Option Pricing, Volatility Skews and Correlation Sensitivity

A difficulty associated with pricing any basket option is deriving an appropriate implied volatility skew. In this work, we consider a simple but market-consistent method commonly used by derivatives practitioners (Deng, 2008; Avellaneda, 2009). In particular, we look to construct implied basket volatility skews from the available index and single stock implied volatility skews.[5]

To do so, let us consider the standard equation for estimating portfolio volatility,

$$\sigma_p^2 = \sum_{i=1}^{N} (w_i \sigma_i)^2 + 2 \sum_{i=1}^{N} \sum_{j>i}^{N} w_i w_j \sigma_i \sigma_j \rho_{ij}. \tag{4.18}$$

Portfolio volatility is thus a function of the weights, volatilities and correlations of the underlying portfolio constituents. If one uses a traded index as the portfolio in question, then the portfolio weights are simply equal to the index weights. Furthermore, if one uses existing implied volatilities as estimates for the respective index and stock volatility parameters, then the only remaining unknowns are the pairwise correlations. Using basic algebra, we can replace these correlations with a single average correlation value and rearrange Equation 4.18 to make this variable the subject of the formula. We are thus able to calculate the average correlation, $\bar{\rho}$, that is *implied* by the current index weights and implied volatilities:

$$\bar{\rho}\,(K,T) = \frac{\hat{\sigma}_I^2\,(K,T) - \sum \left(w_i \hat{\sigma}_i\,(K,T)\right)^2}{2 \sum_{i=1}^{N} \sum_{j>i}^{N} w_i w_j \hat{\sigma}_i\,(K,T)\,\hat{\sigma}_j\,(K,T)}. \tag{4.19}$$

Note that the implied average correlation is a function of strike and term because the underlying implied volatilities are also functions of strike and term.[6]

With an estimate of the average implied correlation surface calculated from available market indices, one can then construct an implied volatility skew for any custom

[5] While more sophisticated (and thus complex) methods do exist for pricing basket options - see Borovkova et al. (2012) and Venkatramanan and Alexander (2011) – the additional accuracy in the calculated volatility skew is not necessary for our illustrative purposes. Furthermore, basket pricing quotes will generally vary far more than for comparative index quotes, making the quest for additional theoretical precision of questionable practical value.

[6] The Chicago Board Options Exchange (CBOE) uses Equation 4.19 with at the money implied volatilities to publish two implied average correlation indices – short-term and long-term – for the S&P 500® Index. According to the CBOE, these indices offer insight into the amount of forward-looking diversification priced into index options relative to individual single stock options.

basket by using an altered version of Equation 4.18 with the given basket weights and known single stock implied volatilities skews. This construction method is market-consistent in the sense that the constructed volatility skew of any index tracking portfolio will always exactly match the traded index volatility skew. In other words, we have essentially calibrated our basket volatility model to match the only basket options that actively trade in the market.

To build intuition on how the price of a basket option changes with implied correlation – this sensitivity is sometimes called rega, or correlation delta – consider an at the money put option written on the current Top40 basket. Figure 4.9 graphs option premium versus implied correlation shifts for four options of varying terms of up to a year. From this, we see that there is a clear and significant positive relationship between option premium and implied correlation and this becomes stronger with term. Premium is also slightly concave across changes in implied correlation, indicating a minor negative second-order effect.

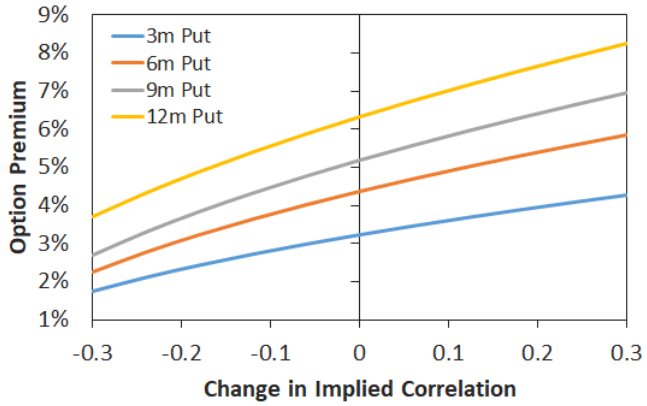


Figure 4.9. At-the-money basket put option premium versus implied correlation shifts.

Although Figure 4.9 gives an indication of how correlation affects a single option strike across multiple terms, it does not consider how correlation sensitivity changes across strikes. To analyse this, we construct a measure of correlation sensitivity by calculating the percentage change in option premium for a 1% shift in implied correlation. Figure 4.10 graphs this correlation sensitivity across moneyness (strike over spot) for option terms of up to a year. We see that correlation sensitivity peaks around the $100 - 110\%$ moneyness range depending on option term and decays as one moves further away from the spot level, particularly on the right-hand side. The increase in correlation sensitivity with respect to term is now also clearly evident.

Having an understanding of how average implied correlation affects basket option pricing provides one with useful intuition on how much diversification potential is being assumed in the basket. Given that the example chosen here reproduces the index skew for zero implied correlation shifts, it therefore also gives an indication of

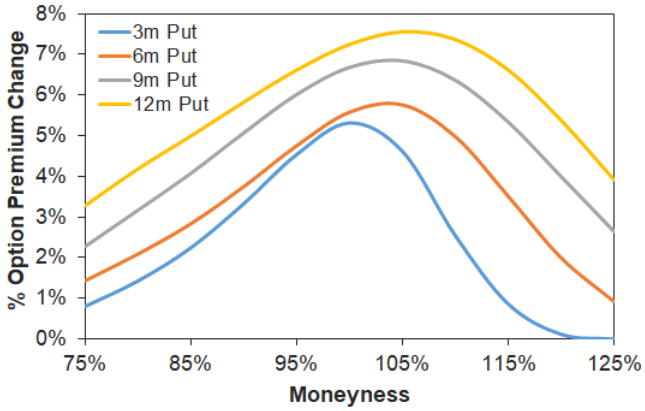


Figure 4.10. Basket put option premium sensitivity to 1% correlation shift across moneyness.

how a basket option's price is likely to differ from readily available index pricing, based on the correlation between that basket and the index.

4.5.2 Introducing Long/Short Basket Options

As discussed throughout this chapter, the most intuitive return decomposition of an active portfolio is into an index component and an active component, which itself can be further decomposed into an active long component and an active short component of equal size. An index hedge removes risk for the index component of the portfolio, leaving only the active long and active short components. The greatest relative risk then faced by the manager is that the gains from the active long component are fully offset by any losses from the active short component. In this case, the most elegant hedging instrument would be one which pays out the difference in value between the active long portfolio and the active short portfolio, thus ensuring a return greater than or equal to the index.

An instrument like this already exists and is called an outperformance option; named as such because one is taking a view on the relative outperformance of one asset versus another (Derman, 1996). In the classical form, the payoff is simply a function of the return differential between two assets. However, if one extends this basic payoff to include a strike level K, then the instrument is referred to as a spread option and has a terminal payoff of $[\omega\,(S_1 - S_2 - K)]^+$ where $\omega = 1$ for a call option and $\omega = -1$ for a put option (Kirk, 1995). Outperformance options can thus be thought of as spread options with a zero strike.

Another way to think of a spread option though is as a long/short basket made up of only two assets, with weights of 100% and -100% respectively. Using this alternative definition, the link between spread options and basket options on the long/short

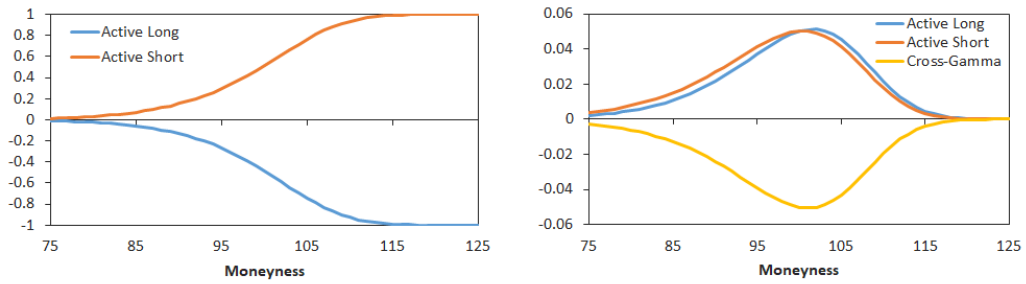
active components becomes clear. If one thinks of the active long and active short components as two separate and singular assets, then a simple method for pricing and understanding the multi-asset long/short active basket option is to rewrite it as a two-asset spread option. A well known closed-form approximation for pricing such European spread options is given by Kirk (1995).[7]

As an illustrative example, consider a 25-stock portfolio with a tracking error of 5% to the Top40. This portfolio has an active share of 48.3%, which is equivalent to the nominal size of the active long and active short portfolios respectively. The active long portfolio consists of 23 stocks and the active short portfolio comprises 19 stocks. In order to use Kirk's (1995) approximation formula, one needs to have implied volatilities for each asset as well as a measure of the correlation between the assets. In our case, implied volatility skews for the long and short portfolios can be constructed using the method discussed in Section 4.5.1 and the implied average correlation skew can be used as a measure of the correlation between the two portfolios. Kirk's (1995) approximation also requires a strike convention for choosing the appropriate implied volatilities for each portfolio. In this work, we use the standard convention that the selected strikes from each portfolio's implied volatility skew should be equal to the spread option's strike. With this in place, we are thus able to price any long/short basket option in a simple and practical manner, while still maintaining the core principle of market-consistent pricing.

Using our example portfolio, Figure 4.11 displays a selection of greeks for a three-month spread option put on the active long/short portfolio and graphed against spread option strike. Studying these greeks and comparing them to more common vanilla option greeks helps one build the necessary intuition about how long/short basket options behave with respect to their underlying asset levels, volatilities and correlations. Note that because the spread option is written on two underlyings, there are two values for each option greek.

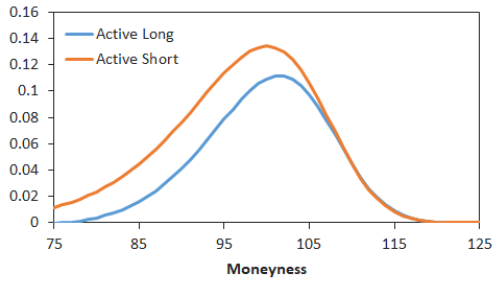
The put option delta values are intuitively negative and positive for the long portfolio and short portfolio respectively, and are very similar to the deltas of vanilla put and call options. The portfolio gammas are both positive and also display a similarity to the vanilla gamma profile, spiking just above the 100 strike level. Note that the longer tail is seen on the left-hand side rather than on the right as for vanilla options. An additional multi-asset greek is cross gamma, measuring the change in delta of one portfolio with respect to a change in the underlying value of the other portfolio. The cross gamma for spread options is negative as one would expect and of similar magnitude to the individual gammas. Vega is larger for the active short portfolio, indicating a higher volatility skew, particularly on the left-hand side. Finally, the

[7] See Venkatramanan and Alexander (2011) for a more accurate – and complex – alternative pricing method.

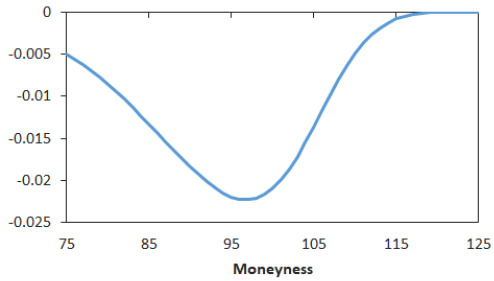


(a) Delta

(b) Gamma and Cross gamma



(c) Vega



(d) Correlation delta

Figure 4.11. Option greeks for a 3-month spread put option written on the active long and active short sub-portfolios of a 5% tracking error portfolio.

correlation delta is approximately an inversion of the three-month profile given in Figure 4.10 for the long-only basket. Again, this is to be expected given that high correlation between the two active portfolios would cause the return spread to decline.

4.6 Conclusion

In this chapter, we have shown how one can approach the question of finding the optimal hedge underlying for any given active portfolio. Starting from the fundamental law of active management, we introduced a framework that allowed us to conduct analysis on simulated realistic active portfolios in order to build intuition as to how hedge mismatch error affects the level of protection afforded by a given hedge. We showed that for typical market conditions, hedge effectiveness declines dramatically when using a hedge portfolio that deviates significantly from the underlying portfolio. This has clear consequences for using generic index options to hedge highly active portfolios.

We also showcased several active return and tracking error decompositions that allow one to precisely quantify and thus manage the sources of risk and reward within a given portfolio. Building on this knowledge, we discussed a mixed integer quadratic

programming formulation that enabled us to search across a large investment universe in order to find the subset of stocks that would most closely replicate a given portfolio's future performance, whilst simultaneously complying with realistic market constraints.

Motivated by our index hedge mismatch findings as well as our success in creating appropriate hedging baskets, we suggested several alternative hedging methods for active portfolios that can provide significantly greater levels of protection than the generic index hedge. From these alternatives, we focussed on long-only basket hedging and long/short basket hedging as a means of creating more appropriate portfolio hedges. At the same time, we discussed a practical, market-consistent pricing methodology for general long/short basket options and provided some initial insight into their behaviour with respect to their underlying asset levels, volatilities and correlations.

There exist several avenues for further research based on the work in this chapter. Firstly, it would be interesting to run a hedging backtest on a set of real-world active portfolios and compare the historical performance and downside risk values when using index options versus custom basket options to hedge the active portfolios. Secondly, this chapter presents a novel method for pricing long/short basket options motivated by practical considerations. It would be useful to rigorously test the accuracy and robustness of this method under a range of assumed market conditions and asset dynamics.

Part III

Chapters in the $\mathbb{P}$-Spectrum

Factor Investing in South Africa

5

> *Your reasoning is excellent, it's only your basic assumptions that are wrong.*
>
> — **Ashleigh Brilliant**
> (Author and cartoonist)

Chapter Synopsis

Risk factors and systematic factor strategies are fast becoming an integral part of the global asset management landscape. In this chapter, we provide an introduction to, and critique of, the factor investing paradigm in a South African setting. We initially discuss the general factor construction process at length and construct a comprehensive range of risk factors for the South African equity market according to international factor modelling standards.

We focus on the size, value, momentum, profitability, investment, low volatility and low beta risk factors respectively. We critically examine the historical behaviour and robustness of these factors, paying particular attention to the issues of long-only versus long/short factors, the impact of size, the effect of rebalancing frequency and data, and the robustness of performance to alternative factor definitions.

We also review how these factors can be used generally in risk management and portfolio management. To this end, we consider factor risk attribution and returns-based style analysis in the risk management space, and multi-factor portfolio construction methods in the portfolio management space.

This chapter is adapted from the journal article by Flint et al. (2017a) and addresses research questions 2b, 3b and 4b as outlined in Section 1.2.3.

5.1 Introduction

Risk factors and the strategies based thereon are fast becoming an integral part of the global asset management landscape.[1] The financial industry has adopted the moniker smart beta to describe such strategies as the term is both highly marketable and sufficiently broad to cover a wide range of investment products. However, in this work we will rather make use of the terms risk factors or risk premia when referring to underlying market drivers, and systematic strategies when referring to the dynamic investment strategies followed in order to gain exposure to underlying risk factors. We do this not only to be more rigorous but also to draw attention to the practical fact that identifying a risk factor and subsequently harvesting returns from that factor are largely separate problems and need to be approached as such.

The latest annual smart beta surveys from FTSE Russell, EDHEC and MSCI all show variations of the same two major trends. Firstly, there are already a number of large international institutional investors that have sizeable factor-based portfolios and secondly, that many more investors are either in the process of reviewing such strategies or are looking to do so in the near future. In order to understand why risk factor investing has shown such a remarkable growth in popularity, it is worth briefly considering the greater history of portfolio management and asset pricing.

Nearly 70 years ago, Markowitz (1952) introduced the efficient frontier approach to asset allocation, which is still the most popular framework for constructing portfolios of assets. Under this framework, an optimal portfolio is defined as the combination of assets that maximises the expected return of the portfolio at a given time horizon for a specified level of portfolio risk (Meucci, 2001). In theory then, the portfolio construction problem had been solved. One simply needed to input the expected returns and covariances of the assets into the framework and out would pop an optimal portfolio specific to one's risk preferences. When applied in practice though, the model was found to be incredibly sensitive to small changes in the estimated mean returns and the optimisation procedure would almost certainly output unreasonable allocations. This behaviour led to Michaud (1989) coining the infamous phrase "error maximiser".

As a result, academics and practitioners alike then focussed their efforts into two separate areas in order to address the framework's weaknesses. The first area was based on all things risk-related: risk-based portfolio construction, more efficient risk estimates, and new risk and diversification measures. The result of this work has culminated in a rich risk budgeting and diversification approach. Roncalli (2013)

[1] The factors and strategies are known by many names, including: risk factors, risk premia, smart beta, alternative beta, systematic strategies, quantitative strategies and rule-based strategies.

provides an excellent review of generalised risk budgeting and Flint et al. (2015b) provides a comprehensive study of diversification in the South African market.

The second area is based on all aspects of creating better expected return estimates. In particular, academics and practitioners went on the hunt for the underlying building blocks of asset classes in a similar manner to the way that physicists have hunted for the increasingly small and elementary particles from which all matter is comprised. The result of this search in the financial industry has given rise to the current factor investing paradigm. Podkaminer (2013) describes risk factors as the "smallest systematic units that influence investment return and risk characteristics" and Cazalet and Roncalli (2014) describe risk factor investing simply as "an attempt to capture systematic risk premia". Homescu (2015) further adds that the aim of factor investing is to construct portfolios in a systematic manner in order to gain exposure to a range of underlying risk factors.

The objective of this research is to construct a comprehensive range of risk factors for the South African equity market, analyse the historical behaviour of these factors and provide an overview of how such factors can be used in risk management and portfolio management. In order to achieve this objective, this research draws heavily on the excellent reviews written by Ang (2014), Cazalet and Roncalli (2014), Amenc et al. (2014), Homescu (2015) and Meucci (2019). We also make reference to Mutswari's (2016) recent work on testing the validity of a number of factor models for South African stock returns.

The rest of this chapter is organised as follows. Section 5.2 reviews the set of linear factor models used in finance and discusses the Fama-French factor models at length. Section 5.3 discusses the general factor construction process and the Fama-French construction methodology in detail. South African risk factors are introduced and thoroughly analysed. Section 5.4 then considers the application of these factors in risk management, focussing on risk attribution and returns-based style analysis. Factor-based portfolio management is discussed in Section 5.5, with emphasis on creating multi-factor portfolios. Section 5.6 concludes and outlines some ideas for further research.

5.2 Linear Factor Models in Finance

Almost all finance studies throughout history have shown that there is a trade-off between risk and return. A natural question for investors then is what level of return can one expect to obtain for exposing oneself to a given level of risk? Traditionally, questions of this nature have been answered by using linear factor models, or

LFMs, which posit a linear relationship between an asset's expected return and its covariance with the risk factors incorporated in the model.

Meucci (2019) states that LFMs are used in almost every step of the risk and portfolio management process, including asset pricing, risk attribution and modelling, alpha prediction, portfolio optimisation and asset allocation. LFMs are also the cornerstone of factor investing as they are the main quantitative tool used to create systematic factor strategies. In this section, we briefly review the key LFMs used in the asset pricing literature and discuss at length the commonly used Fama-French-type factor models.

5.2.1 CAPM and APT

The capital asset pricing model (CAPM) was introduced by Sharpe (1964) and serves as the basis for all other factor models of asset returns. Based on the framework defined by Markowitz (1952), Sharpe showed that the risk premium on an asset (or portfolio of assets) was a linear function of a single market risk premium, represented by the market-capitalisation index. Mathematically, the CAPM states that

$$\mathbb{E}\left[R_i\right] - R_f = \beta_i \left(\mathbb{E}\left[R_m\right] - R_f\right), \tag{5.1}$$

where R_i and R_m are the returns on the i^{th} asset and market portfolio respectively, R_f is the risk-free rate, $\mathbb{E}\left[\cdot\right]$ represents the expectation and β_i is the beta – or sensitivity – of the i^{th} asset to the market portfolio, calculated as the ratio of the covariance of the asset and the market portfolio to the variance of the market portfolio:

$$\beta_i = \frac{\mathbb{C}ov\left[R_i, R_m\right]}{\mathbb{V}ar\left[R_m\right]}. \tag{5.2}$$

Beta thus measures the level of non-diversifiable, systematic risk embedded within any asset. Given that there is only a single market risk factor, CAPM states that the reward for taking on additional risk is directly proportional to the underlying market risk. Therefore, everyone should hold the market portfolio in equilibrium as it is the only risk that is truly rewarded. While extremely elegant, there have been countless studies since its introduction that have shown that the theoretical CAPM is not validated by empirical evidence.

Ross (1976) proposed an alternative model, known as arbitrage pricing theory (APT) based on the increasing evidence of multiple market risk premia. Ross posited that

the return of an asset is driven by a combination of random market factors and that this can be modelled with an LFM:

$$R_i = \alpha_i + \sum_{j=1}^{J} \beta_i^j \mathcal{F}_j + \varepsilon_i, \tag{5.3}$$

where α_i is a constant, β_i^j is the sensitivity of asset i to factor j, $\mathcal{F}_j$ is the return on factor j, and ϵ_i is the independent and identically distributed (*iid*) stock-specific error term, which is also independent from any of the risk factors. It can be shown from Equation 5.3 that under APT, the risk premium on an asset is given by

$$\mathbb{E}\left[R_i\right] - R_f = \sum_{j=1}^{J} \beta_i^j \left(\mathbb{E}\left[\mathcal{F}_j\right] - R_f\right). \tag{5.4}$$

Equations 5.3 and 5.4 form the basis of nearly all risk attribution systems and systematic factor strategies. One of the challenges in using the APT though is that it is left to the user to define what the underlying market risk factors really are. In this vein, Cazalet and Roncalli (2014) define three main risk factor categories. The first category comprises factors based purely on statistical asset data – e.g. principal components analysis risk factors. The second category comprises factors based on macroeconomic data – e.g. inflation and GDP growth. The final category comprises factors based on market data. This can be further classified into those factors based on accounting data – e.g. size and value – and those based on price data – e.g. momentum and low volatility. In this work, we focus mostly on the third category of risk factors.

5.2.2 The Fama-French Model and its Extensions

Fama-French Three-Factor Model

Based on the prior empirical studies that analysed numerous potential risk factors, Fama and French (1993) proposed a three-factor model for equity stock returns, which has since become the industry standard. This model linearly combines accounting- and price-based factors in the form

$$R_i - R_f = \alpha + \beta_i^m \left(R_m - R_f\right) + \beta_i^{smb} R_{smb} + \beta_i^{hml} R_{hml} + \varepsilon_i. \tag{5.5}$$

R_{smb} is the return on a long/short portfolio of small/big market capitalisation stocks and R_{hml} is the return on a long/short portfolio of high/low book-to-market stocks.[2] These are known as the size factor and value factor respectively. Because market capitalisation and value ratio indicators are correlated, Fama and French (1993) use

[2] Factor construction is discussed at length in Section 5.3.

a two-way sorting procedure to strip out any confounding factor effects. The value factor thus captures the value premium that is independent of the effect of size and the size factor captures the size premium that is independent of the effect of value.

There has been much literature aimed at assessing the appropriateness of the Fama-French three-factor model in equity markets worldwide. In the South African context, van Rensburg (2001) and van Rensburg and Robertson (2003) provide some of the earliest comprehensive assessments of Fama-French based APT models on the Johannesburg Stock Exchange (JSE). Although not testing the exact Fama-French three-factor model, they show convincingly that one needs to incorporate several risk factors in order to accurately model the cross-section of equity returns on the JSE. More recent studies in the same vein include the works of Mutooni and Muller (2007), Basiewicz and Auret (2009, 2010), Strugnell et al. (2011) and Muller and Ward (2013), among others. Although these studies report differences in the magnitudes and significance levels of certain equity risk factors, they all conclude that a broader APT-based factor model is required to model South African equity markets correctly. The difference in study results is also to be expected, given the variations in data period and method across the various studies. As both Amenc et al. (2014) and Cazalet and Roncalli (2014) note, risk factors can be both cyclical and market-specific.

Carhart Four-Factor Model

Motivated by the evidence provided by Jegadeesh and Titman (1993) on the existence of significant medium-term price momentum trends, Carhart (1997) introduced a four-factor model based on Fama and French's work but including a momentum factor. This has since become the standard model used in fund performance and persistence literature. Mathematically, the Carhart four-factor model is given as

$$R_i - R_f = \alpha + \beta_i^m \left(R_m - R_f\right) + \beta_i^{smb} R_{smb} + \beta_i^{hml} R_{hml} + \beta_i^{wml} R_{wml} + \varepsilon_i, \quad (5.6)$$

where R_{wml} represents the return on a long/short portfolio of winner/loser stocks, based on the previous 12-month's price performance. Although initially met with severe scepticism, the momentum factor is now referred to as the "premier market anomaly" (Fama & French, 2008). Studies have confirmed the presence of this anomaly across numerous geographies and asset classes, making it the most prevalent market factor to date (Moskowitz et al., 2012; Asness et al., 2013). Perhaps the reason for this pervasiveness is because the momentum factor is in essence a behavioural artefact, driven by cognitive biases which are unlikely to disappear in the

near future (Antonacci, 2013). The same is perhaps not true about the justifications of the size and value factors.

Fama-French Five-Factor Model

In the time since Fama and French's (1993) initial work, many authors have shown that the three-factor model and even the four-factor model may well not be sufficient to explain the variation in the cross section of asset returns. To this effect, Fama and French (2015) introduced a novel five-factor model which included factors relating to the profitability and level of investment made by a company. In contrast to their original model, which is based on APT and empirical market research, the justification for the five-factor model stems from the bottom-up dividend discount model. Specifically, Fama and French (2015) suggest that expected stock return, as modelled by the dividend discount model, is based on three variables, namely the book-to-market ratio, expected earnings and expected growth in book equity – what they dub 'investment'. From their investigations, they posit the following five-factor model:

$$R_i - R_f = \alpha + \beta_i^m (R_m - R_f) + \beta_i^{smb} R_{smb} + \beta_i^{hml} R_{hml} + \beta_i^{cma} R_{cma} + \beta_i^{rmw} R_{rmw} + \varepsilon_i, \tag{5.7}$$

where R_{cma} represents the return on a long/short portfolio of conservatively/aggressively invested stocks, and R_{rmw} represents the return on a long/short portfolio of robust/weak profitability stocks. Apart from the dividend discount model, the inclusion of these two factors was also influenced by the work of Novy-Marx (2013) and others, who showed that high profitability (or quality) stocks are rewarded with a significant and consistent premium, even after accounting for the return stemming from the original risk factors. Asness et al. (2014) have since refined Novy-Marx's proxy of profitability/quality and proposed a new long/short factor of quality/junk stocks, where quality is defined as a composite score based on the dividend discount model and comprising numerous single accounting values. For the remainder of this paper, we will focus only on Fama and French's (2015) version of the profitability (i.e. quality) factor.

Asness et al. Six-Factor Model

Given that the Fama-French five-factor model is motivated by the dividend discount model, which describes the long-term behaviour of expected stock returns, the absence of the shorter-term momentum factor becomes somewhat more understandable. However, its exclusion is still surprising given that these very same authors

named momentum as the premier market anomaly. In addition to this observation, Asness et al. (2015) also suggest that value and momentum are complementary risk factors and should be placed together. As a result, they propose a six-factor model extension which includes the momentum factor and makes use of a slightly adjusted value factor:

$$R_i - R_f = \alpha + \beta_i^m (R_m - R_f) + \beta_i^{smb} R_{smb} + \beta_i^{hml} R_{hml}^* + \beta_i^{wml} R_{wml} + \beta_i^{cma} R_{cma} + \beta_i^{rmw} R_{rmw} + \varepsilon_i. \tag{5.8}$$

According to their results, the six-factor model provides a more complete explanation of the variation in historical US stock returns than the five-factor model and the adjusted value factor, which was shown to be nearly redundant by Fama and French (2015) before adjustment, now remains a significant risk factor.

Other Risk Factors

In what has now become one of the classic empirical finance papers, Harvey et al. (2015) surveyed hundreds of asset pricing papers published over the last fifty years and tallied more than 300 factors that are purported to explain the variation in the cross-section of expected returns. This concerted exercise in data mining led to Cochrane (2011) coining the phrase “the factor zoo”.

The proliferation of purported factors is also partly a consequence of the popularity of the factor investing paradigm: factors are now everywhere and everything has become a factor. Cazalet and Roncalli (2014) suggest that this is arguably the most pernicious fantasy in the factor investing literature. Instead, they state that there are only a handful of risk factors that represent true risk premia or market anomalies. Ang (2014) suggests four main criteria for determining whether an observed market phenomenon is actually a true risk factor:

1. It should have strong support in academic and practitioner research and strong economic justifications.

2. It should have exhibited significant premiums to date that are expected to persist.

3. It should have history available during both quiet and turbulent market regimes.

4. It should be implementable in liquid, traded instruments.

Although the final criterion is not strictly required if only using the factor model in a risk attribution setting, it is still vitally important for creating tradable systematic factor strategies.

The factors we have discussed so far are all considered to be true risk factors in the sense that they are prevalent across nearly all markets studied to date, have valid economic and/or behavioural justifications, and have histories stretching back more than a hundred years in some cases. In addition to these well-established risk factors, there are also a handful of recently discovered factors that are fast becoming accepted as true risk factors.

Two such recent factors attempt to capture the observed empirical phenomena that low volatility stocks outperform high volatility stocks and, similarly, that low beta stocks outperform high beta stocks. Ang et al. (2006) and Blitz and van Vliet (2007) popularised the idea of the low volatility factor and showed significant premium levels attached to this factor across a range of markets. Baker et al. (2014) and Frazzini and Pedersen (2014) among others have since confirmed their results and refined the economic rationale, further justifying the observed risk premia.

The low beta factor can be traced all the way back to Black (1972) and the leverage effect. Despite this lengthy history, the factor has only come back into vogue in the last ten years. Interestingly, van Rensburg and Robertson (2003) showed early on that the low beta anomaly commanded a significant premium in the South African equity market and could be accessed by sorting portfolios into quintiles based on their CAPM betas.

Other common factors not considered in this work are the carry (i.e. dividend yield), liquidity and quality factors. The carry risk factor is perhaps the most easily accepted in South African markets, where both the FTSE/JSE Dividend Plus Index and dividend-based unit trusts have existed for many years already. The liquidity factor is also easily appreciated in South African markets given its extremely high levels of concentration and the constant problem of capacity that many of the larger fund managers are faced with. Even though the strategy is accessed by going long illiquid stocks and shorting liquid stocks, it is unlikely that one could ever easily trade a South African liquidity factor in any decent size. For this reason, we leave this factor for future consideration. Finally, we have the quality factor. As mentioned above, the Fama and French (2015) profitability factor is essentially equivalent to the Novy-Marx (2013) version of quality. Although the more involved definition by Asness et al. (2014) is arguably a better proxy for the true quality factor, it is also considerably more complicated to manufacture. For the sake of simplicity then, we leave this more advanced quality factor for future consideration.

5.3 South African Equity Risk Factors

In Section 5.2, we outlined several of the most popular APT-based factor models used in practice which have become essential risk and portfolio management tools. Although the selection of an optimal model specification remains an open question, it is clear that the underlying risk factors used in these competing models will continue to remain relevant for the foreseeable future. To this end, there are several online, open-source risk factor databases for large international equity markets.[3] However, and despite the South African-based factor studies mentioned earlier, a similar database does not exist – or at least is not publicly available – for the South African equity market.

One of the goals of this research is to create a growing database of South African equity risk factors constructed as per the international asset pricing literature. In particular, we construct seven Fama-French style factors (size, value, momentum, profitability, investment, low volatility and low beta) and several factor-sorted portfolios and publish these in an open-source, online factor data library.[4] Our hope in doing so is to make an independent factor database available to industry and academia that enables them to run a number of risk and portfolio management factor applications in line with international best practice.

5.3.1 Generalised Factor and Signal Processing

The factors discussed in this work are based on the Fama-French portfolio sorting methodology, which we will outline shortly. However, it is important to realise this is simply a special case of a more general signal processing framework. Meucci (2019) outlines three steps in the general allocation policy for systematic strategies. Firstly, process the set of current information into one or more factor signals. Secondly, transform these signals into a single set of consistent characteristics (i.e. expected return estimates) on the underlying stocks. Thirdly, construct optimal portfolio weights as a function of the transformed signal characteristics.

The initial step can be broken further into data collection, signal generation and signal processing. Consider a momentum signal for example. After collecting the requisite price data and correcting for any corporate actions and dividend payments, one uses a defined function to create factor scores. This could be as simple as prior 12-month return or something more complicated like a Hull moving average filter.

[3] For example, see the comprehensive risk factor databases maintained by Kenneth French (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html) and Andrea Frazzini (http://www.econ.yale.edu/~af227/data_library.htm).

[4] The online South African factor data library can be accessed at https://legaeperesec.co.za/.

Finally, these scores are filtered over time and/or cross-sectionally in order to create factor signals. Common filtering techniques include smoothing over time, scoring to reduce volatility, ranking cross-sectionally, twisting ranks nonlinearly, and trimming or Winsorising outliers.

The second step is not usually carried out when constructing single factors but is vitally important when considering multiple factors. For example, consider a universe of stocks that have both momentum and value scores. One then needs to define a methodology for creating a single consistent characteristic value for each stock that is consistent with both sets of factor scores. Such methods can vary from basic portfolio sorts to complex nonlinear programming solutions. We revisit this point in Section 5.5.1.

Finally, create an optimal portfolio based on the estimated stock characteristics, a given satisfaction index and a set of constraints. This implementation step is ultimately what separates systematic factor strategies from underlying risk factor portfolios. In special cases, one can directly trade the underlying risk factors but usually investors are faced with real-world constraints that make this impossible. For example, long-only investors wanting to gain exposure to the long/short Fama-French value factor need to use optimisation techniques in order to maximise targeted factor exposure while minimising unwanted factor exposures. See Section 5.5 for more on this.

5.3.2 Constructing South African Risk Factors

We now consider the Fama-French construction methodology in light of the general factor framework outlined above. The data set consists of the 383 constituents of the FTSE/JSE All Share Index (ALSI) over the period January 1996 to August 2016. All available total return and fundamental stock data were obtained from Bloomberg and INet for the 20-year period. Due to severe limitations on available fundamental data, the initial starting date had to be moved forward to December 2002, thus yielding a final sample period of just less than 14 years.

The majority of Fama-French risk factors are based on fundamental stock variables, with the remainder based on price information variables. The definitions of each such variable were kept consistent with the relevant international literature. At any particular month in the analysis window, the factor variables are defined as follows:

- *Size* is defined as the market value of the stock as at the end of the previous month. The shares in issue are taken directly from the underlying FTSE/JSE

index data and multiplied by the index-recorded share price to obtain the gross market capitalisation.

- *Value* is defined as the ratio of book value to market value (BtM). This ratio is computed by taking the most recent book value six months prior to the current month and dividing it by the market value as at the end of the previous month. This is slightly different to the original definition but is in line with the alteration proposed by Asness and Frazzini (2013).

- *Momentum* is defined as the prior twelve month total stock return, less the prior month's return to account for any short-term reversal effects.

- *Profitability* is defined as the ratio of operating profit (total annual revenue, net of sales and other expenses) to the most recent book value for the previous year.

- *Investment* is defined as the relative growth in total assets six months prior to the current month.

- *Low volatility* is defined as the standard deviation of weekly total stock returns measured over the three years prior to the current month. If three years of weekly return data are not available, a smaller history is used with the minimum period required being one year. This is the factor definition proposed by Blitz and van Vliet (2007).

- *Low beta* is defined as the CAPM beta estimated from weekly excess total stock returns and excess ALSI returns, measured over the three years prior to the current month. If three years of weekly return data are not available, a smaller sample is used with the minimum period required being one year. This is the factor definition proposed by Blitz and van Vliet (2007).

The stock universe available for factor construction at any given month is taken as the historical ALSI constituent basket for that month. In order to isolate the true premia of the underlying factors, Fama and French (1993) employ a basic two-way portfolio sorting methodology. We create long/short factor returns in a consistent manner:

1. First rank all stocks according to their size score. Using the 50^{th} percentile as a break point, create two subsets of stocks, namely *Big* (all the stocks above the break point) and *Small* (stocks below the break point).

2. Independently rank all the stocks according to their value score. Taking the 30^{th} and 70^{th} percentiles as break points, construct three value subsets; namely, *High* value above the 30^{th} percentile, *Neutral* value between the 30^{th} and 70^{th}, and *Low* value (i.e. growth) stocks below the 30^{th} percentile.

3. Repeat the previous step to construct stock subsets on the basis of momentum, profitability, investment, low volatility and low beta scores respectively. Note that in the case of investment, low volatility and low beta, the portfolio below the 30th percentile is the one which is expected to render the positive return.

4. Use the two-way size/factor sort in order to create equally-weighted factor portfolios, as depicted in Table 5.1. For example, the size/value sorting procedure gives one six portfolios: namely, Small Value, Small Neutral and Small Growth, and Big Value, Big Neutral and Big Growth.

5. Construct long/short factor returns by averaging the returns on the Small High and Big High factor portfolios and subtracting the average of the returns on the Small Low and Big Low factor portfolios. Repeat this for each set of sorting tables to create the six size-agnostic factor portfolios.

6. Construct long/short size factor returns for each of the independent two-way sorting tables by averaging the returns on the Small High, Small Neutral and Small Low factor portfolios and subtracting the average of the returns on the Big High, Big Neutral and Big Low factor portfolios. The final long/short size factor return is then calculated as the average of the various size factor returns across all factors included in the model.

Table 5.1. Depiction of the two-way factor portfolio sorts for the Carhart four-factor model.

Book-to-Market Value Portfolios			
	Growth	*Neutral*	*Value*
Small	SG	SN	SV
Big	BG	BN	BV

12-1m Momentum Portfolios			
	Losers	*Neutral*	*Winners*
Small	SL	SN	SW
Big	BL	BN	BW

Following Step 5 above, the long/short value factor return is calculated as

$$R_{hml} = \frac{1}{2}\left[R\left(SV\right) + R\left(BV\right)\right] - \frac{1}{2}\left[R\left(SG\right) + R\left(BG\right)\right] \tag{5.9}$$

$$= R^{+}_{hml} - R^{-}_{hml} \tag{5.10}$$

$$= \frac{1}{2}\left[R\left(SV\right) - R\left(SG\right)\right] + \frac{1}{2}\left[R\left(BV\right) - R\left(BG\right)\right] \tag{5.11}$$

$$= \frac{1}{2}\left(R^{small}_{hml} - R^{big}_{hml}\right). \tag{5.12}$$

Equations 5.10 and 5.12 show how to decompose the long/short factor return into separate long and short components as well as into separate size components. These decompositions also represent perhaps the two most common constraints faced by investors in the risk factor space: namely, long-only and capacity constraints. We will revisit this in Section 5.3.3.

Following Step 6, the size factor return from the size/value portfolios is calculated as

$$R_{smb}^{val} = \frac{1}{3}\left[R\left(SV\right) + R\left(SN\right) + R\left(SG\right)\right] - \frac{1}{3}\left[R\left(BV\right) + R\left(BN\right) + R\left(BG\right)\right]. \quad (5.13)$$

A similar calculation is done for the return on the size/momentum portfolios, R_{smb}^{mom}, and the final size factor return is thus given as

$$R_{smb} = \frac{1}{2}\left(R_{smb}^{val} + R_{smb}^{val}\right). \quad (5.14)$$

One departure from the methodology of Fama and French is the continued use of two-way rather than n-way sorts for the larger factor models. We do this because of the discrepancy between the size of the South African stock universe, which ranges from 150 to 171 stocks over the 14 year period, and the size of the US stock universe, which numbers in the thousands. Even if one were to use only two portfolios per factor, a four-way sort would cause the average portfolio size to drop to only ten stocks. This is clearly not large enough to ensure a well-diversified portfolio free from stock-specific risk.

Rebalancing of the value, profitability and investment factors occurs annually at each December-end. The low volatility and low beta factors are rebalanced quarterly, beginning from December-end, and the momentum factor is rebalanced monthly. As noted in Step 4, the standard methodology is to create equally-weighted factor portfolios, although one can also consider cap-weighted portfolios. If any constituents of the factor portfolios delist during the holding period, an appropriate portfolio rebalance is done as at the close on the day prior to delisting as per standard indexing rules.

In summary, the process outlined above ensures that we create realistic and tradable daily risk factor returns over the complete sample period. Finally, we use the ALSI total return less the three-month NCD rate as a proxy for the excess market factor.

5.3.3 Factor Analysis

Figure 5.1 displays the cumulative log-performance of the eight South African long/short risk factors over the full 14-year sample period. Equal-weighted factors

are represented by the solid lines and cap-weighted factors by the dashed lines. The most striking observation is that the scale of the momentum factor is significantly larger than any of the other factors, including the (excess) market factor. Apart from the international evidence that suggests that momentum generally does command the largest risk premium (Antonacci, 2013), the strong performance is likely also due to the underlying equity market's strong performance over the sample period, coupled with the extreme level of concentration. On average, the ten largest stocks in the ALSI have historically accounted for nearly 60% of the total index value (Flint et al., 2013a). Therefore, any strong underlying equity market trend – positive or negative – is almost certainly driven by this handful of large counters. Such a feature is exactly what the momentum factor attempts to capture. Lastly, one must also remember that the momentum portfolio rebalances monthly and thus a large proportion of this return could be lost in practice due to high turnover costs.

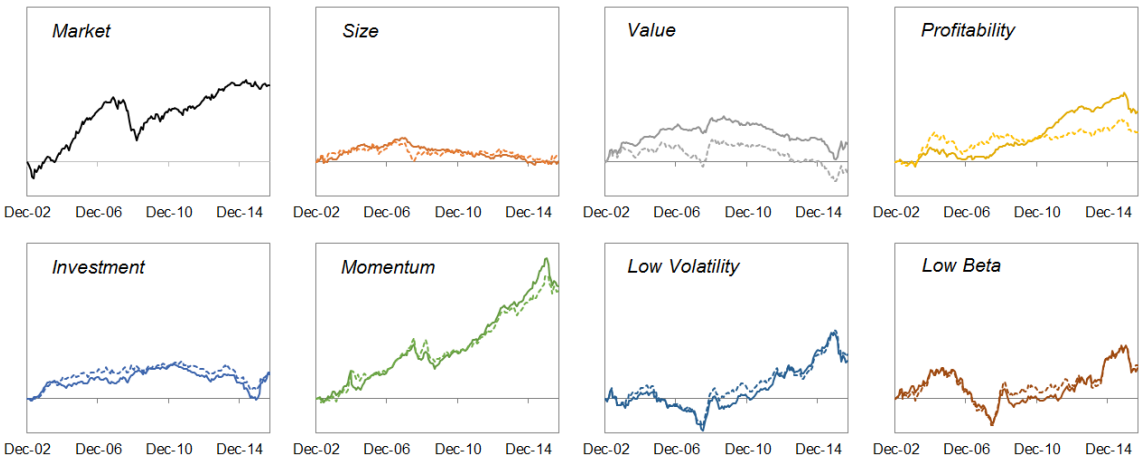


Figure 5.1. Cumulative log-performance of equal-weight (solid) and cap-weight (dashed) South African risk factors, Dec 2002 to Aug 2016.

Figure 5.1 also shows that the weighting scheme used in the Fama-French sorting procedure can impact the performance of the risk factor, although the magnitude of the effect is very factor-dependent. The discrepancy in equal- and cap-weighted factors is most obvious for the value and profitability factors. The remaining factors though show little to no difference in either trend or return magnitude.

Over the complete period, the size premium has remained consistently small and has in fact been slightly negative since the 2008 financial crisis; in line with the findings of Strugnell et al. (2011). As Table 2 shows, the expected return on the size factor is only 0.1%, a stark contrast to the 12.4% return on the momentum factor. The value factor, arguably the most well-known and accepted risk premium, has also struggled since the financial crisis, thus giving only a 2% annual return over the full period. This perhaps explains the poor performance of many South African value funds over the last decade.

We also note that the investment factor has not been particularly well rewarded over the last five years, showing a similar contraction as in the value premium. This is perhaps somewhat understandable as the level of annual asset growth and the book value of a company are surely somewhat connected on a fundamental level. This hypothesis is also supported by the fact that investment is the only factor to show a positive correlation of 0.31 to value, albeit still small in absolute terms.

In contrast to the size, value and investment factors, profitability has shown strong performance over the last decade, particularly over the financial crisis and recovery period. This makes intuitive sense though, as this factor essentially proxies the quality of a company's earning streams and one would expect high quality earnings streams to have been the least affected by the crisis, and also to have participated strongly in the subsequent recovery rally. It also supports the recent industry trend in international markets of focussing on quality-sorted versions of the other factors (Gray & Carlisle, 2014; Gray & Vogel, 2016).

Table 5.2. Equal-weight long/short factor summary statistics, Dec 2002 to Aug 2016.

	Market	Size	Value	Profitability	Investment	Momentum	Low Vol.	Low Beta
CAGR	8.66%	0.11%	1.98%	5.70%	2.68%	12.43%	4.25%	3.38%
Volatility	15.95%	7.21%	10.42%	9.21%	10.18%	14.86%	15.90%	18.04%
Kurtosis	0.55	0.36	1.66	3.14	5.66	2.11	1.67	0.22
Skewness	-0.12	-0.14	0.23	-0.93	1.01	-0.60	-0.22	-0.13
Min. Return	-14.25%	-7.63%	-11.08%	-12.90%	-7.68%	-17.63%	-17.11%	-13.71%
Max. Return	13.05%	4.61%	9.46%	6.53%	16.68%	11.96%	16.13%	16.94%
Sharpe Ratio	0.54	-1.01	-0.52	-0.18	-0.46	0.34	-0.20	-0.22
Max DD	-47.4%	-32.8%	-47.6%	-26.6%	-40.6%	-28.8%	-45.4%	-56.2%

Tables 5.2 and 5.3 also highlight some interesting points about the low volatility and low beta factors. In contrast to what one might expect, Table 5.2 shows that these two factors have the second highest and highest return volatility respectively. However, this phenomenon actually confirms the rationale motivating these factors; namely that there is an inverse relationship between volatility or beta and the actual risk premium awarded to the stock. Whatever the economic reasoning though, we note that both factors have performed strongly since the financial crisis. The strong positive correlation of 0.78 between the returns of these two factors suggests that they are capturing overlapping parts of the same underlying factor, which one would expect. However, we do note a higher kurtosis, lower volatility and lower maximum drawdown attached to the low volatility factor. One final point of interest with these factors is their strong positive correlations of 0.62 and 0.55 respectively to the profitability factor. We leave this observation for future research.

Table 5.3. Equal-weight factor correlation matrix and correlations between equal-weight and cap-weight factors, Dec 2002 to Aug 2016.

Equal-Weight	Market	Size	Value	Profitability	Investment	Momentum	Low Vol.	Low Beta
Market	1.00							
Size	-0.37	1.00						
Value	-0.20	0.09	1.00					
Profitability	-0.14	0.08	-0.26	1.00				
Investment	0.01	-0.04	0.31	-0.51	1.00			
Momentum	-0.01	-0.04	-0.45	0.42	-0.75	1.00		
Low Volatility	-0.49	0.20	-0.03	0.62	-0.46	0.33	1.00	
Low Beta	-0.48	0.20	0.02	0.55	-0.33	0.35	0.78	1.00

EW vs. CW	Market	Size	Value	Profitability	Investment	Momentum	Low Vol.	Low Beta
	0.94	0.73	0.84	0.68	0.70	0.74	0.95	0.92

As with all asset classes, risk factors also display varying degrees of cyclical behaviour. Although this is graphically evident in Figure 5.1, we provide more tangible evidence of this feature in Table 5.4, which presents factor statistics for three contiguous sub-periods of 412 years. In particular, we consider the bull market from December 2002 to June 2007, the crisis and recovery rally from June 2007 to December 2011, and the positive but slowing market from December 2011 to August 2016. These are referred to as periods 1, 2 and 3 in the table.

Table 5.4. Long/short factor performance across three sub-periods.

Statistic	**Period**	**Market**	**Size**	**Value**	**Profit.**	**Investment**	**Momentum**	**Low Vol.**	**Low Beta**
	1	21.64%	7.59%	11.40%	1.88%	6.86%	13.41%	-3.50%	-0.91%
CAGR	2	-2.52%	-4.47%	0.78%	11.57%	3.29%	7.25%	7.45%	2.27%
	3	8.26%	-2.29%	-5.24%	3.93%	-1.76%	18.38%	9.06%	8.81%
	1	16.13%	7.89%	9.50%	9.75%	7.70%	15.52%	13.49%	17.44%
Volatility	2	19.61%	6.68%	8.98%	7.18%	8.56%	12.30%	15.69%	16.64%
	3	10.67%	6.67%	12.16%	10.31%	13.29%	16.63%	18.13%	20.00%
	1	1.34	0.02	0.42	-0.57	-0.07	-0.07	0.39	-0.48
Sharpe	2	-0.13	-1.78	-0.74	0.58	-0.48	-0.01	-0.01	-0.31
Ratio	3	0.77	-1.45	-1.04	-0.34	-0.69	0.66	0.09	0.07
	1	-21.3%	-8.0%	-16.4%	-9.8%	-24.9%	-26.8%	-26.8%	-39.3%
Max. DD	2	-47.4%	-15.2%	-15.5%	-9.5%	-31.1%	-45.4%	-45.4%	-56.2%
	3	-15.4%	-47.6%	-26.6%	-40.6%	-34.2%	-35.5%	-35.5%	-38.1%

There are meaningful differences in nearly all factors and statistics across the three periods. In particular, we see that the largest drawdowns for most of the factors occurred in the most recent period and specifically over the last two years. Two of the main reasons for this – although certainly not the only ones – are that the proportion

of South African-specific risk to global risk in the local market has been increasing since 2012 (Flint et al., 2015b), and that some of the largest ALSI constituents have recently experienced significant company-specific events. This highlights the general need to ensure that one is effectively diversified against those risks which do not carry any discernible risk premia as well as being diversified across the risk factors that do carry a positive premium over the long-term. It is this last reason that has driven the rise of multi-factor portfolios, discussed further in Section 5.5.

5.3.4 Factor Robustness

As with any empirical financial study, one needs to address the question of robustness. In particular, one should always be cognisant of the fact that the constructed factor portfolios will always only be noisy proxies of the true underlying risk factors. To this end, we consider the robustness of such factors to the choices made during the construction process. We have already highlighted one such choice in Figure 5.1 by showing the effect that weighting scheme can have. In this section we scrutinise a number of other important construction choices.

Long-only versus Long/Short Factors

One of the most pertinent constraints for many investors is the inability to short sell assets either at all or to the extent that they would wish. This raises the issue of whether long-only factor proxies are able to provide similar risk factor exposure in comparison to their long/short counterparts. A fundamental challenge in factor investing is the investability of the underlying factor portfolios. It is all well and good to create theoretically appealing long/short factor portfolios and use these for risk attribution – see Section 5.4.1 – but this may all be for nought if one cannot effectively allocate capital to such portfolios. Hence the proposal of long-only factor portfolios. Although such portfolios will contain residual market risk by construction, we believe that their interpretation as risk factors still remains valid. Furthermore, given that all the factors will on average have similar levels of market risk exposure, this residual risk should largely cancel out in any risk attribution exercises.

Figure 5.2 compares the performance of the long-only component of each factor (solid lines) against the complete long/short portfolios (dashed lines), and Table 5.5 gives the long-only factor summary statistics. In the case of the market factor, we are comparing the absolute market return with its excess-to-cash counterpart. There is a stark contrast in performance between all the long-only and long/short portfolios. It is also clear that the long-only risk factors – barring size – comfortably outperform the absolute market return.

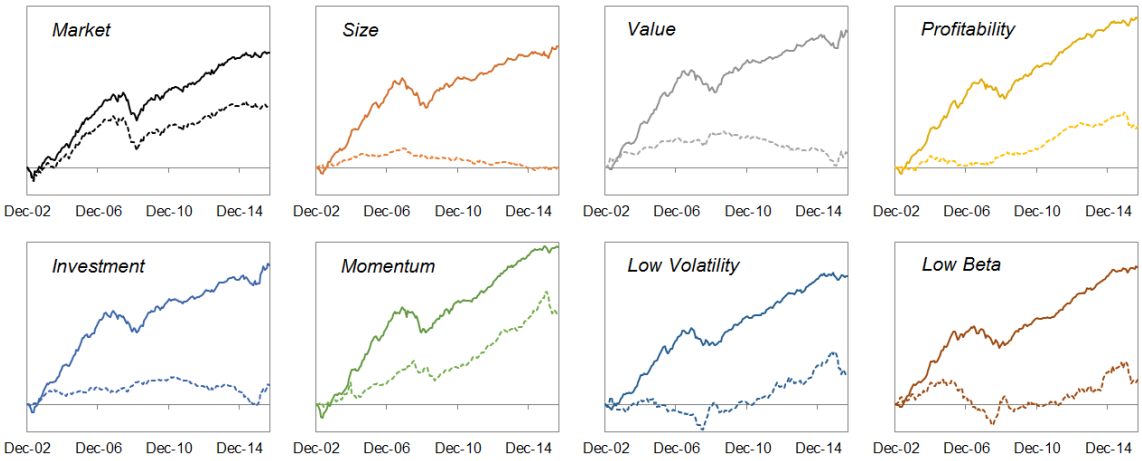


Figure 5.2. Cumulative log-performance of long-only (solid) and long/short (dashed) South African risk factors, Dec 2002 to Aug 2016.

Table 5.6 gives the correlation matrix of the long-only factors as well as the correlations between the long-only and long/short versions of each factor. The supposition of contaminating latent market exposure is proven by the strong positive correlations with the market factor. Furthermore, the correlations between each risk factor are now also very high as a result. Considering the correlations between long-only and long/short factor versions, it is interesting to note that despite the similarity in trend between the two momentum factors, the correlation between these two factors is only mildly positive at 0.29. This serves as a poignant reminder about the pitfalls of conflating price trend and return correlation. What Figure 5.2 does suggest though is that the short component of the momentum factor provides only limited benefit across the period.

Table 5.5. Equal-weight long-only factor summary statistics, Dec 2002 to Aug 2016.

	Market	Size	Value	Profitability	Investment	Momentum	Low Vol.	Low Beta
CAGR	16.92%	17.85%	20.22%	22.49%	20.66%	23.74%	18.97%	20.40%
Volatility	15.32%	13.62%	15.19%	13.51%	15.51%	15.32%	11.98%	12.74%
Kurtosis	0.34	1.31	0.60	0.53	0.14	1.41	1.47	1.35
Skewness	-0.18	-0.73	-0.20	-0.42	0.01	-0.58	-0.70	-0.59
Min. Return	-13.10%	-14.07%	-13.14%	-11.87%	-10.01%	-14.27%	-13.07%	-11.58%
Max. Return	12.99%	8.26%	11.99%	10.07%	16.89%	12.47%	8.71%	10.28%
Sharpe Ratio	1.10	0.77	0.84	1.12	-0.86	1.07	0.97	1.02
Max DD	-40.4%	-42.2%	-35.1%	-30.8%	-33.3%	-37.5%	-27.5%	-32.5%

Factor Size Effects

Another constraint faced by many investors is that of capacity. Even if one has the ability to short, it may be that the majority of a factor's return stems from the Small sub-portfolios of the factor. Such a size bias would imply limited investment capacity

Table 5.6. Long-only factor correlation matrix and correlations between long-only and long/short factors, Dec 2002 to Aug 2016.

Long-Only	Market	Size	Value	Profitability	Investment	Momentum	Low Vol.	Low Beta
Market	1.00							
Size	0.71	1.00						
Value	0.66	0.90	1.00					
Profitability	0.76	0.91	0.84	1.00				
Investment	0.75	0.88	0.90	0.81	1.00			
Momentum	0.80	0.87	0.76	0.91	0.80	1.00		
Low Volatility	0.62	0.86	0.80	0.89	0.69	0.81	1.00	
Low Beta	0.55	0.81	0.74	0.83	0.66	0.78	0.85	1.00

L-O vs. L/S	Market	Size	Value	Profitability	Investment	Momentum	Low Vol.	Low Beta
	n.a.	0.18	0.41	0.17	0.42	0.29	0.07	0.16

owing to the small market capitalisation of the underlying stocks and potential illiquidity issues. Several authors have suggested that such factor size biases exist in many markets (Homescu, 2015). If present in the highly concentrated South African equity market, this bias would have serious ramifications on the prospect of large-scale South African factor investing. Figure 5.3 breaks down each factor return into its Big (solid line) and Small (dashed line) sub-portfolios as per Equation 5.12. Note that these sub-portfolios are still long/short combinations and hence are of similar magnitudes to the complete factor returns shown in Figure 5.1.

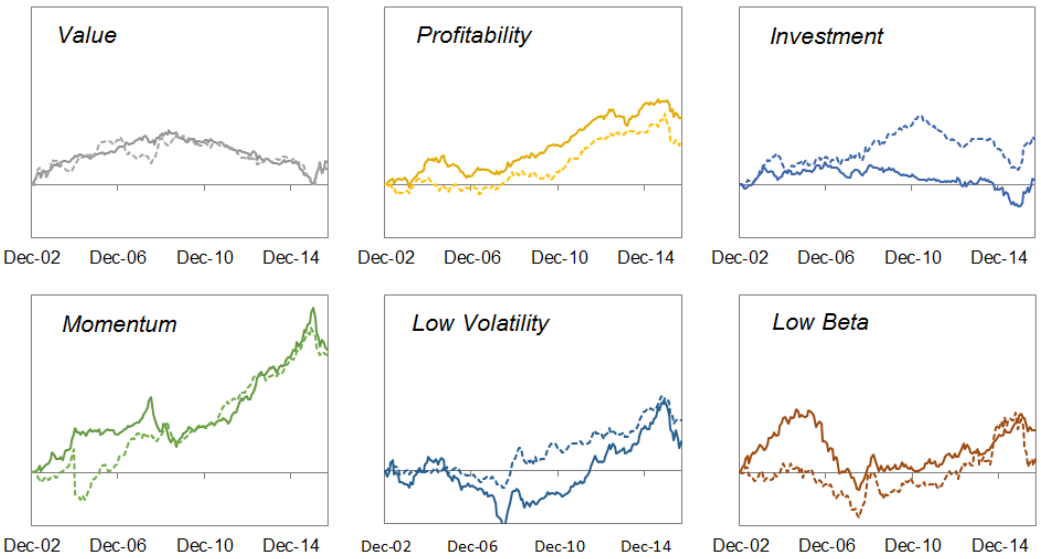


Figure 5.3. Cumulative log-performance of big (solid) and small (dashed) South African risk factors, Dec 2002 to Aug 2016.

Momentum and value don't display any significant size bias. Of the remaining four, profitability displays a small, persistent bias towards large stocks, while investment displays a persistent bias towards small stocks. Low volatility and low beta display discrepancies between big and small long/short portfolios that vary over the sample period.

Rebalance Frequency and Date

Value, profitability and investment portfolios are rebalanced annually at the beginning of each year. Low volatility and low beta portfolios are rebalanced quarterly with the first rebalance occurring at the beginning of the year, and momentum portfolios are rebalanced at the end of each month. The choice of rebalance frequency for each factor is driven by the time frame over which the factor signal decays. There is also the more practical issue that any benefit gained from more frequent rebalancing may be offset by the additional transaction costs. For the majority of our factors, the time frame of the risk premia is well established. However, given the relatively new 'discovery' of the low volatility and low beta factors, the effect of rebalance frequency is less well documented. To this end, we compared the returns from the low volatility and low beta factors when rebalancing monthly, quarterly, biannually and annually, finding only minor differences.

Another rebalancing issue to consider for those factors with longer holding periods is the choice of month in which to enact the rebalance. As above, we test how much of an impact moving rebalance dates has by considering the returns from twelve value factors each rebalanced in different months of the year and again find no significant return differences. Although it may seem odd to include such a non-result in our research, it is an incredibly important one from a practical implementation perspective. Furthermore, it showcases the fact that the factor construction methodology outlined in Section 5.3.2 is generally robust to rebalancing choices.

Factor Sorting Extremity

The standard Fama-French two-way sorting procedure uses the 50^{th} percentile of the size score and the 30^{th} and 70^{th} percentiles of the factor scores as the relevant sorting break points. A natural question then is whether using more extreme percentile break points results in larger factor risk premia. The trade-off here is that one essentially creates 'purer' factor portfolios but at the cost of increasing the portfolio's

idiosyncratic risk. This is particularly pressing in the South African equity market, which only contains around 170 counters in the tradable ALSI universe.

To test the robustness of the factors to the sorting methodology, we create extreme factor portfolios using the 20^{th} and 80^{th} percentiles of the relevant factor scores as sorting break points. Figure 5.4 gives the comparison between the standard (solid line) and extreme (dashed line) factors. Somewhat surprisingly, only the extreme value and momentum factors show any significant difference to their standard counterparts. In both cases, the divergence of the extreme factor performance is most evident in the last ten years and seems to be linked to outperformance during periods of financial stress. We leave further investigation of this phenomenon for future research.

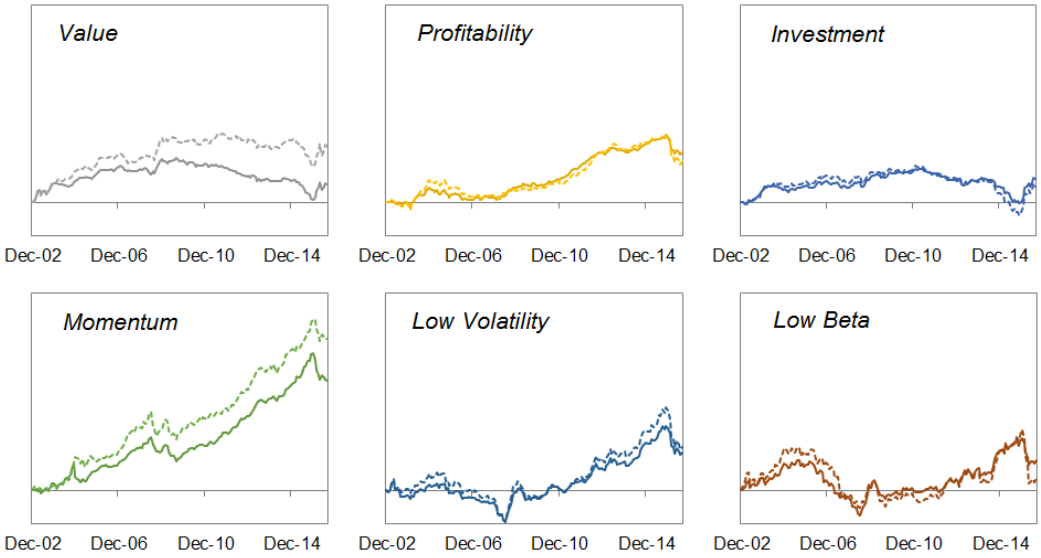
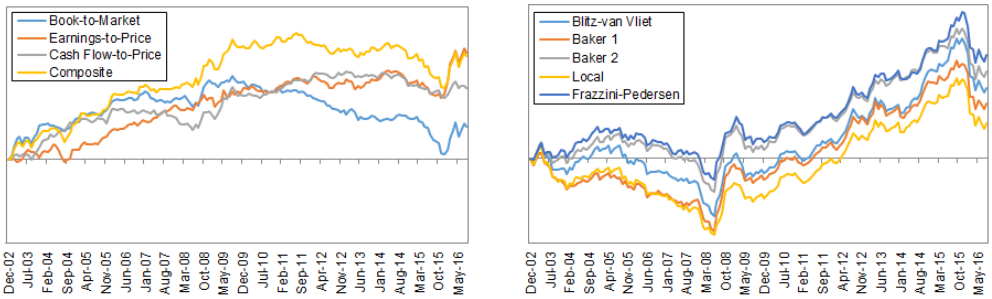


Figure 5.4. Cumulative log-performance of standard (solid) and extreme (dashed) South African risk factors, Dec 2002 to Aug 2016.

Alternative Factor Definitions

Although varying the choice of sorting percentile can in some respects be considered as using an alternative factor definition, the more obvious alternative is to use a different fundamental stock characteristic as a proxy for the underlying factor score. As an example, we have already discussed the multiple definitions of the quality factor in Section 5.2.2. In a similar vein, a number of authors have considered alternative measures for value and for low volatility. Popular alternative value score candidates include earnings-to-price, cash flow-to-price and a composite score based on these two metrics as well as the original book-to-market ratio (Amenc et al., 2014). In the low volatility literature, the alternatives are not different risk measures but rather different calculation methods for volatility; the main variables being the



(a) Value factor variants **(b)** Low volatility factor variants

Figure 5.5. Cumulative log-performance of variant factor definitions, Dec 2002 to Aug 2016.

length of the historical estimation window and the frequency of return data.[5] Blitz and van Vliet (2007) suggest using three years of weekly data, Baker et al. (2014) suggests using either 60 monthly or 60 weekly return observations, local industry research considers three years of monthly data, while Frazzini and Pedersen (2014) suggest one year of daily return data.

Figure 5.5 gives the cumulative log-performance of long/short factors based on these alternative value and low volatility scores. The variant return range for both factors is fairly substantial and particularly so for the value factor. Furthermore, the behaviour of the variant value factors differs significantly throughout the period, which suggests that the selected stock characteristics capture different aspects of the true value risk factor. The relative outperformance of the composite value score supports this suggestion and also highlights the importance of reducing signal noise; in this case achieved by averaging out the characteristic-specific noise.

For the low volatility factor, performance of the factors all show the same pattern, indicative of the fact that only the calculation method is changing, rather than the measure itself. Interestingly, both of the top performing variants are those that use the smallest estimation window – 1 year and 60 weeks respectively – as well as higher frequency data – daily and weekly respectively.

5.4 Factor-Based Risk Management

At its core, portfolio management is about making decisions: when to buy or sell any given asset and in what quantity. These decisions are made in order to add value to a passive benchmark, be it a nominated index or a cash-based rate. In this setting, 'adding value' is usually defined in two ways. The first is by achieving a positive

[5] Similar calculation method alternatives apply to the beta factor score.

return, or alpha, over and above the nominated benchmark at an acceptable level of risk.[6] The second is by achieving a specified target return at a lower level of risk than that of comparable passive market products.

In both cases, the strength of any portfolio decision should be measured by how much value it generates for the fund, conditional on the market and fund constraints faced by the manager over the performance period. Flint et al. (2015a) show how one can use the fundamental law of active management (FLOAM) framework of Clarke et al. (2002) in order to decompose a fund's relative return and risk into contributions from each of the underlying fund constituents. We build on this work here but consider instead the idea of risk attribution rather than risk decomposition. In particular, we consider how to attribute a fund's risk – absolute or relative – to a given set of external risk factors. Such an attribution lets one identify what kinds of factor risk a fund is exposed to and furthermore calculate how large these factor bets are. Knowing this allows one to make informed and efficient investment decisions.

5.4.1 Factor Risk Attribution and the Factor Efficiency Ratio

Given a series of fund returns – absolute or relative – we can use one of the LFMs described in Section 5.2.2 to attribute risk to the underlying risk factors constructed in Section 5.3. Although more difficult than attributing risk to the fund's constituents, Meucci (2007, 2019) describes how one can still attribute fund risk to a set of external risk factors in an additive fashion. Furthermore, if one does have sight of the fund's holdings, it is possible to attribute risk similarly for each of the underlying constituents so that the fund's factor risk contributions can be written as a linear combination of the constituents' factor risk contributions (Roncalli & Weisang, 2016). This is perhaps the most important factor application in the risk management space. Consider the pedagogical example below.

Table 5.7. Simulated fund risk factor exposures.

	Fund 1	Fund 2	Fund 3	Fund 4	Benchmark
Market	0.5	0.1	0.2	0.2	0.25
Size	0.2	0.5	0.1	0.2	0.25
Value	0.2	0.2	0.5	0.1	0.25
Momentum	0.1	0.2	0.2	0.5	0.25

We select the Carhart four-factor risk model and make use of long-only risk factors. Let us assume that there are four funds that are currently under investigation. We simulate monthly returns for these funds using the factor exposures given in Table

[6] All portfolio management should be considered benchmark-relative, even if the selected benchmark is a constant value of zero.

5.7. A small random alpha term (centred at 0.25%) and a larger random noise term (centred at zero) are added to each fund's monthly return.

Table 5.8 gives a comprehensive factor risk attribution for both the absolute and relative risk of each fund based on the Carhart four-factor model. By construction, the estimated betas are very similar to the input fund exposures and the R^2 of the risk model is very high. Table 5.8 also shows the risk contributions of each factor as well as the catch-all residual term. These values are also closely related to the estimated beta levels owing to the high correlation between the risk factors as well as their similar volatility levels. Finally, contributions to tracking error are also calculated across the funds for each risk factor. Because of the good fit of the risk model, most of the tracking error stems from the fund-specific noise term.

Table 5.8. Carhart model factor risk attribution.

	Fund 1	**Fund 2**	**Fund 3**	**Fund 4**	**Benchmark**
R^2	94.2%	95.5%	94.6%	95.4%	95.9%
Volatility	14.27%	13.88%	13.01%	14.68%	13.86%
Tracking Error	5.37%	4.71%	4.89%	4.48%	n.a.
			Betas		
Alpha	0.38%	0.10%	0.35%	0.18%	0.00%
Market	0.52	0.05	0.14	0.23	0.23
Size	0.29	0.48	0.11	0.24	0.26
Value	0.15	0.24	0.51	0.06	0.24
Momentum	0.05	0.22	0.15	0.49	0.25
			Risk Contributions		
Market	52.9%	4.0%	13.8%	21.3%	22.8%
Size	23.8%	44.9%	10.2%	20.2%	24.0%
Value	12.9%	24.3%	55.5%	5.4%	23.4%
Momentum	4.6%	22.3%	15.1%	48.4%	25.7%
Residual	5.8%	4.5%	5.4%	4.6%	4.1%
			Tracking Error Contributions		
Market	30.9%	26.9%	14.3%	-0.4%	n.a.
Size	-0.8%	9.5%	12.7%	-0.7%	n.a.
Value	3.9%	0.1%	-1.6%	4.7%	n.a.
Momentum	4.6%	0.8%	15.8%	29.5%	n.a.
Residual	61.3%	62.6%	58.8%	66.8%	n.a.

In the context of factor investing, where investors are actively seeking exposure to the underlying risk factors, risk and tracking error contributions become incredibly important as they provide a means of quantifying and thus evaluating such exposure. To this end, Hunstad and Dekhayser (2015) introduce the factor efficiency ratio (FER) as a means of gauging the amount of intended versus unintended factor risk

exposure in a given fund (or asset). Letting $\mathcal{F}_d$ represent the set of K desired factors, we can write

$$FER\left(\mathcal{F}_d\right) = \frac{\sum_{i=1}^{K} RC_i}{1 - \sum_{i=1}^{K} RC_i}, \tag{5.15}$$

where RC_i is the generic risk contribution stemming from the i^{th} desired risk factor. Hunstad and Dekhayser originally consider the contributions to active risk (i.e. tracking error) but one can just as easily use any convex risk measure to calculate risk contributions.[7] The FER is interpreted as follows: for every $X\%$ of risk stemming from the desired factor set, the fund takes on an additional 1% of risk from undesired factors. Therefore, the higher the FER, the more efficient the fund is at gaining desired factor exposure.

Consider again the four fund example outlined in Table 5.8 and further assume that all of these funds are marketed as composite value/momentum indices. Using this as our desired factor set, we calculate FERs of 0.21, 0.87, 2.40 and 1.17 for each of the funds respectively. Based on these scores, it is clear that Fund 3 provides one with the most efficient exposure to the desired value and momentum factors.

5.4.2 Return-Based Style Analysis and Fund Replication

Sharpe (1992) introduced the concept of returns-based style analysis (RBSA), which is used extensively in the fund management literature. In essence, RBSA is a form of constrained regression that allows one to draw inference on funds for which only historical return data is available. Sharpe suggested using factors based on asset classes and interpreted the model output as being indicative of a manager's style mix. Ultimately, given a set of historical fund returns, RBSA estimates the static mix of tradable market indices or factors that most closely replicates the fund's returns, R_{pt}. Letting β represent the vector of factor exposures, we can formulate the RBSA estimation problem as follows:

$$\begin{aligned} &\underset{\beta}{\operatorname{argmin}} \sum_{t=1}^{T} \left(R_{pt} - \sum_{j=1}^{J} \beta_j \mathcal{F}_{jt} \right)^2 \\ &s.t. \quad \beta_j \geq 0 \\ &\qquad \sum \beta_j = 1. \end{aligned} \tag{5.16}$$

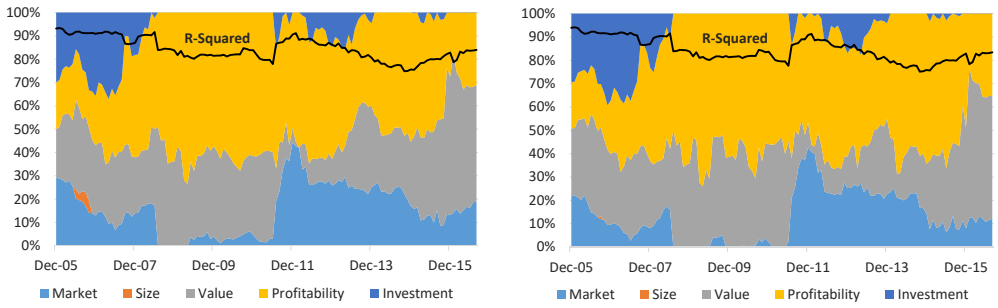
In a sense, the RBSA betas represent the long-only weights of the replicating style portfolio. However, this is not strictly true because the betas remain fixed across the estimation window whereas portfolio weights would change in line with the

[7] Note that one has to treat negative risk contributions with caution when calculating the FER as they can materially change its interpretation. The simplest solution is to take absolute values of all risk contributions and replace the '1' in the denominator with the sum of the absolute risk contributions.

performance of the underlying factors. Several improvements to the initial RBSA methodology have been suggested to address this, and other issues. These include the use of the Kalman filter, corrections for heteroscedasticity and the inclusion of structural break detection mechanisms. Another point which is common to all regression but generally not considered in RBSA is that of confidence intervals around the estimated betas.[8] For example, a style weight of 30% with a confidence interval of +/- 2% should be viewed very differently to a weight of 30% with a confidence interval of +/- 20%.

A variation of RBSA that is particularly relevant in the index tracking space is to solve for the initial number of 'shares' (rather than betas) of each factor that minimises the tracking error (rather than sum of squared errors) of the estimated style portfolio to the given fund returns. Therefore, one can not only use the RBSA framework to measure a given fund manager's style mix but also, after some adjustment, to create tradable replicating portfolios for a fund. This alternative usage has been explored at length in connection with hedge fund replication.

In a similar vein to Section 5.4.1, we showcase RBSA with an illustrative example. We attempt to uncover the style mix of the FTSE/JSE Dividend Plus Index by making use of the long-only Fama-French five-factor model. Figure 5.6 displays the RBSA factor exposures (left panel) and the adjusted-RBSA replicating weights (right panel) from December 2005 onward. We fit the models on a monthly basis using rolling 36-month windows and record the static betas and end-of-period weights respectively.



(a) RBSA betas **(b)** RBSA replicating weights

Figure 5.6. RBSA betas and end-of-period weights for the FTSE/JSE Dividend Plus Index and the long-only Fama-French five-factor model, Dec 2005 to Aug 2016.

Although the RBSA betas are similar to the replicating weights, one can still easily see the discrepancies in Figure 5.6. The R^2 of both models is consistently high, meaning that the majority of variation in the index is well-captured by the five-factor model.

[8] See Lobosco and diBartolomeo (1997) for an approximation formula for constructing confidence intervals around the constrained betas.

The style mix of the index varies considerably over the period, which suggests that the dividend yield measure is actually a composite signal for a number of underlying risk factors. The largest exposure over the period has been to the profitability factor – in line with the yield-driven nature of the index – with the remainder mostly split between the value and market factors. Investment exposure is sporadic and has been absent over the last three years. Size is irrelevant for the Dividend Plus index, which is to be expected given that the index is limited to large-cap and mid-cap stocks.

Table 5.9 gives the RBSA betas and end-of-period weights for the 36-month period ending at 31 August 2016. Although similar in nature, there is still an absolute difference of 13.7% across the factors. This difference is driven by the varying performance of the underlying factors and is directly related to the level of factor dispersion over the period.

Table 5.9. RBSA betas and replicating weights for the FTSE/JSE Dividend Plus Index, 31 Aug 2016.

	Market	Size	Value	Profitability	Investment
RBSA Betas	19%	0%	50%	31%	0%
95% CI	(7.8%,30.3)	(-11.9%,11.9%)	(40.3%,59.7%)	(21.5%,40.5%)	(-10.7%,10.7%)
RBSA Weights	12.2%	0%	52.6%	35.3%	0%
Diff. to Betas	-6.9%	0%	2.6%	4.3%	0%

5.5 Factor-Based Portfolio Management

In addition to the risk management applications given above, risk factors are also used extensively in portfolio management. And while the concept of factor investing is definitely not new, the recent rise of the smart beta phenomenon has attracted significant attention to this area.

In the last several years, the focus has started to move away from identifying additional risk factors and towards constructing optimal multi-factor portfolios. While some authors have said that there is no formal framework in place for combining systematic factor strategies (De Franco et al., 2016), the fact of the matter is that the majority of the existing optimisation frameworks – risk/return or risk-only – are fully capable of incorporating both factor portfolios and factor-based risk/return views. Furthermore, the allocation policy for systematic strategies outlined by Meucci (2019) provides one with a fully general framework for creating optimal multi-factor portfolios in the presence of transaction costs and fund constraints.

In this section we discuss several ideas on how to create such multi-factor portfolios, ranging from the very simple to the fairly complex. Note that most of these are based on concepts that we have already introduced and analysed in preceding sections.

5.5.1 Portfolio Mixing and Integrated Scoring

According to Fitzgibbon et al. (2017), two of the most common approaches for creating multi-factor portfolios are the *portfolio mix* and *integrated score* methods. Portfolio mixing is simply the linear combination of factor portfolios constructed from single-variable sorting procedures. For example, consider a value portfolio based solely on the top quintile of book-to-market stocks and a momentum portfolio based solely on the top quintile of twelve-month return stocks. These portfolios would then be taken as existing building blocks and the only challenge facing the investor would be to set an appropriate weight for each portfolio. Viewed in this light, portfolio mixing can be thought of in a similar manner to the decisions made in strategic asset allocation.

The integrated score approach goes one step further by mixing the underlying factor scores ex ante rather than mixing given factor portfolios ex post. The Fama-French two-way sorting methodology – whereby stocks are selected based on their respective factor score ranks relative to a set of constant percentile break points – is perhaps the simplest example of the integrated score approach. In general, the integrated score approach combines individual factor scores in some manner to create a single, unified score. Figure 5.7 displays this concept graphically and confirms that the field of (non)linear programming provides investors with a natural set of tools for creating optimal integrated multi-factor scores, and thus optimal multi-factor portfolios.

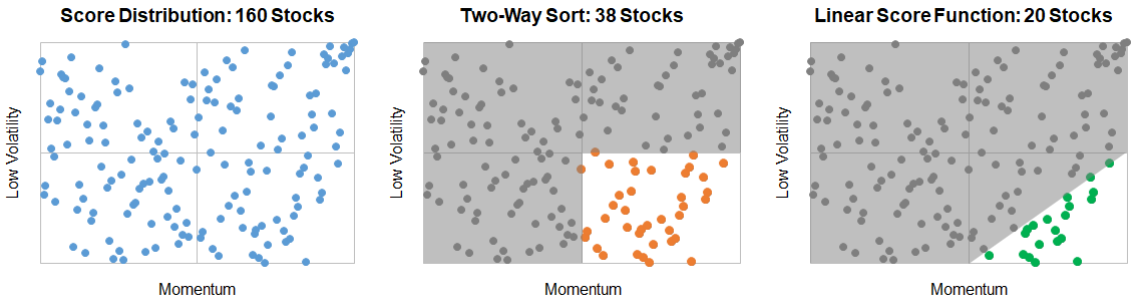


Figure 5.7. Integrated scoring examples for momentum and low volatility.

Lastly and very importantly, Hoffstein (2016) points out that one needs to consider the speed of factor decay when creating these integrated signals. This is particularly relevant when combining the fast-decaying momentum signal with slower signals like value or profitability, for example.

5.5.2 Constrained Factor Optimisation

A more technically rigorous approach than those given above is to view the construction of an efficient multi-factor portfolio as a constrained optimisation problem. Although more complex, this approach allows an investor to construct a multi-factor portfolio that is as consistent with their return objectives and risk preferences as their constraint set will permit. There are a number of optimisation frameworks available to investors, including classical mean-variance and risk-based investing (Richard & Roncalli, 2015), among others.[9] Below we sketch out two candidate optimisation approaches that could be used to create constrained optimal multi-factor portfolios.

The first approach makes use of the risk attribution framework introduced in Section 5.4.1. Assuming that one is given a risk factor model, the problem then becomes finding the underlying stock weights that provide the requisite exposure to the targeted risk factors, whilst minimising undesired factor exposures. If exposure is defined in terms of beta, then one needs to solve for the portfolio of assets that minimises the total distance between estimated and targeted betas, where the target levels for the undesired factors are set to zero. Alternatively, if exposure is defined in terms of risk contributions, then there are two options available. The first option is similar to the beta optimisation but where one instead specifies target risk contribution levels. The second option is to solve for the portfolio of assets that maximises the FER for the set of desired factors. FER optimisation is arguably more intuitive and will likely provide more robust results due to the fact that it simultaneously accounts for the desired and undesired factor exposures in a single monotonic metric. Of the two approaches, we therefore favour FER maximisation.

The second optimisation approach makes use of mixed integer programming (MIP). A mixed integer program is one in which some variables are continuous and some are integers. Such a setting is ideal for problems in which one has to first select a subset of assets from the available universe – the integer variables – and subsequently search for the set of weights – the continuous variables – that minimises an objective function under a set of constraints. In general, mixed integer programs can be quite hard to solve unless one can formulate the problem in a very particular way. Thankfully, we are able to set up both linear (MILP) and quadratic (MIQP) mixed integer programs for most portfolio construction problems which can be solved fairly easily – albeit slowly – with freely available optimisation toolboxes and heuristic solvers.[10]

[9] Please see Homescu (2014) for a comprehensive review of the available portfolio construction frameworks.

[10] For example, Flint et al. (2015a) successfully use the MIQP approach to replicate the Top40 index with only a small number of stocks and construct optimal hedging baskets for active portfolios.

One of the main issues with multi-factor investing is smoothly transitioning between risk and return preferences in the factor space to risk and return preferences in the asset space. This is not a trivial exercise. One way of linking the factor and asset spaces in a manner which does not add additional estimation error would be to combine the integrated score approach with the risk attribution optimisation by means of an MILP. Figure 5.8 presents an example of this combined approach for a low volatility and momentum multi-factor portfolio using scoring data as at August 2016.

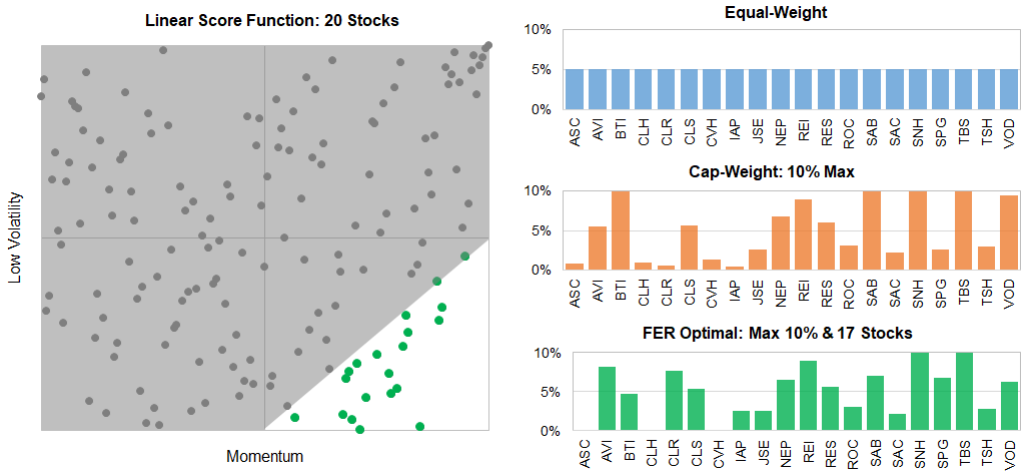


Figure 5.8. Multi-factor portfolio construction with Integrated Scoring and MILP optimisation of the factor efficiency ratio.

Firstly, one uses the integrated score as a screening tool to find the subset of assets that display the fundamental factor characteristics most in line with the desired factor set. Taking this subset of factor-screened assets as an input, one then solves the MILP problem for the maximum FER portfolio under the given constraints, where the choice of assets included in the portfolio and the subsequent weights attached to the chosen assets are both variables in the optimisation. Introducing the integrated score screen and subsequently maximising the portfolio's FER obviates the need to explicitly assign factor-consistent expected return estimates to each asset – a difficult task – and thus also reduces the potential for estimation error in the optimisation.

5.6 Conclusion

Risk factors and systematic factor strategies are fast becoming an integral part of the global asset management landscape. In this research, we have attempted to provide an introduction to, and critique of, the factor investing paradigm in a South African setting.

We created a range of long/short and long-only risk factors for the South African equity market according to the standard Fama-French factor construction methodology; namely size, value, momentum, profitability, investment, low volatility and low beta. Historical risk and return characteristics were shown to vary significantly across the factors as well as across market regimes. Momentum has been the most rewarded factor historically. Low volatility, profitability and low beta have also shown positive risk premia, while the size factor seems to be non-existent in South Africa. We then tested factor robustness at length and showed the effect that each of the major decisions taken in the factor construction process can have. The largest such effect stems naturally from the choice of long-only or long/short factors. Interestingly, we found that, barring size, all long-only factors handily outperformed the market.

In addition to constructing this factor database, we also showcased several risk factor applications. In the risk management space, we considered risk attribution to factors and introduced the factor efficiency ratio as a measure of how efficiently a fund gained exposure to a set of desired risk factors. We also considered returns-based style analysis with long-only risk factors and showed how this could be used to estimate a manager's style mix or to create a replicating factor portfolio.

In the portfolio management space, we considered the issue of creating multi-factor portfolios. We discussed simple approaches such as portfolio mixing and integrated scoring, and more complex approaches based on solving for target risk contributions or optimising the factor efficiency ratio for the desired factors. Finally, we introduced the mixed integer programming framework as a means of combining the integrated scoring approach with the risk attribution optimisation approach in a robust manner, thus allowing one to smoothly transition between preferences and constraints in the non-tradable factor space and the tradable asset space.

There exist numerous avenues for further research based on the work in this chapter. Given that this research is by definition an introduction of the factor investing paradigm in the South African equity market, each of the topics discussed here can, and should, be examined in more detail. One example of such is the practical issue of risk factor construction and strategy implementation, an incredibly important topic for industry practitioners. There are a number of questions within this area that require proper examination. For example, how much of an effect can the portfolio weighting scheme have on the performance of single factor and multi-factor portfolios?

6

Regime-Based Tactical Allocation for Factor and Balanced Portfolios

> *Why should I change with the times, when the times are obviously wrong?*
>
> **— Ashleigh Brilliant**
> (Author and cartoonist)

Chapter Synopsis

It is now an accepted fact that the majority of financial markets worldwide are neither normal nor constant, and South Africa is no exception. One idea that can be used to understand such markets that has been gaining popularity recently is that of regimes and regime-switching models.

In this chapter, we consider whether regimes can add value to the asset allocation process. Four methods for regime identification – economic cycle variables, fundamental valuation metrics, technical market indicators and statistical regime-switching models – are discussed and tested on two asset universes – long-only South African equity factor returns and representative balanced portfolio asset class returns.

We find several promising regime indicators and use these to create two regime-based tactical allocation frameworks. Out-of-sample testing on both the equity factor and balanced asset class data shows very promising results, with both regime-based tactical strategies outperforming their respective static benchmarks on an absolute and risk-adjusted return basis.

This chapter is adapted from the conference proceedings by Flint et al. (2017b) and addresses research questions 2c and 3c as outlined in Section 1.2.3.

6.1 Introduction

It is now an accepted fact that the majority of financial markets worldwide are neither normal nor constant, and South Africa is no exception. Flint et al. (2012; 2014) examined the statistical properties of South African equity index returns and highlighted the following key points:

- Daily index returns display volatility clustering and a strong negative correlation to volatility.
- Market returns are negatively skewed and fat-tailed for return horizons up to one year.
- Asset class returns and fund returns are asymmetric in both volatility and correlation across various market states.
- Return distributions and particularly average return estimates change significantly depending on the specified historical period.
- Extreme outlier returns – both positive and negative – occur more frequently than one might expect and can have a significant impact on long-term portfolio returns.

Researchers have found similar behaviour in a wide variety of markets worldwide, to the point where many of these return characteristics are now termed stylised facts (Cont, 2001). As a result, practitioners and academics alike have turned to new frameworks and models that are consistent with these observations. In the derivative space, this manifests in the form of curved implied volatility surfaces and stochastic or local volatility models for the valuation of exotics options (Seymour, 2011). In the portfolio management space, this results in extensions of the Modern Portfolio Theory (MPT) framework for time-dependent asset dynamics like GARCH volatility and dynamic conditional correlation, or tail risk measures such as value-at-risk and expected shortfall (Seymour et al., 2015). One framework that has been gaining popularity recently in both of these areas is that of regimes and regime-switching models.

Ang and Timmermann (2012) identify three reasons for the popularity of the regime-based framework. Firstly, regimes are intuitive and can naturally fit the ex-post narratives that investors use to explain market moves. Estimated regimes are often found to tie up with low and high volatility periods, up- and down-trending return periods and/or changes in underlying macroeconomic policy. Secondly, regime models are capable of accurately capturing the nonlinear and non-normal

stylised facts outlined above. Figure 6.1 depicts this by showcasing a mixture of two normal distributions. Notice the significant negative skew and fat tails of the mixed distribution. Thirdly, because regime models are generally constructed as a linear combination of (log)normal distributions, they are simple to understand and implement. A regime-based framework thus allows one to capture the intricacies of the market while still affording analytical tractability and familiarity.

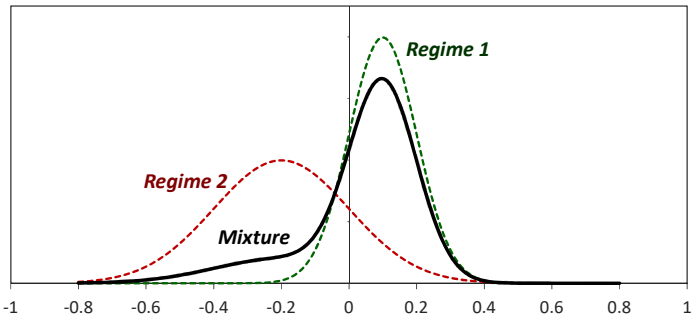


Figure 6.1. Mixture of two normal distributions.

Although South African research on the subject is limited, there have been some studies that incorporate regimes into the investment process. For example, Flint et al. (2014) examine the use of regimes in a hedge timing process and Seymour et al. (2016) consider the use of regimes in portfolio optimisation, both finding promising results. In this research, we extend this work by considering the use of a regime-based framework to enhance the asset allocation process. If one assumes that markets oscillate between divergent regimes, then it stands to reason that an asset allocation process that tactically changes portfolio exposure to account for these regime changes should add value relative to a static portfolio mix. The goal of this work is thus to examine whether such a regime-based asset allocation process can indeed add value in the South African context.

Throughout this research, we make use of two specific asset universes. The first universe consists of seven South African long-only fundamental equity factors; namely, size, value, profitability, investment, momentum, low volatility and low beta. These factors are constructed according to international industry standards but using South African stock data. Please see Chapter 5 for a complete outline of the factor construction process. Factor return data for the period January 2003 to March 2017 was downloaded from the open-source Legae Peresec factor data library.[1] We select the long-only return database constructed from a constrained All Share Index (ALSI) stock universe as a representation of a tradable set of equity factors. The second universe represents the set of asset classes most commonly found in a South African balanced portfolio; namely, local equity (FTSE/JSE All Share Index), local bonds (FTSE/JSE All Bond Index), local property (FTSE/JSE SA Property Index), global

[1]The Legae Peresec factor data library can be accessed at https://legaeperesec.co.za/.

equity (FTSE World Index), global bonds (JP Morgan Government Bond Index), global commodities (RJ CRB Commodity Index) and the USDZAR exchange rate.

The rest of this chapter is organised as follows. Section 6.2 motivates the use of regimes in finance and discusses and implements four of the most popular regime identification methods and variables. Section 6.3 then builds on this regime classification work and tests the out-of-sample performance of two regime-based asset allocation frameworks. The first framework is based on technical indicators and is implemented on the equity factor universe and the second is based on a regime-switching model of financial turbulence and is implemented on the balanced portfolio universe. Section 6.4 concludes and outlines some ideas for further research.

6.2 Identifying Market Regimes

Below, we outline some of the most popular methods and variables currently used for defining and estimating economic environments and market regimes. Note that some of the methods described here are not strictly regime-switching models in a statistical sense. However, they still have the goal of categorising the market into different underlying states and are commonly used in practice.

6.2.1 Macroeconomic Environments: Yield Spread, Inflation and CLI

The idea of the economic or business cycle dates back to the early 1800s and was formalised in 1860 by the French economist, Clement Juglar. Juglar argued that economic prosperity oscillated in some systematic fashion around a long-term trend and that a full cycle was likely to be between 7 to 11 years. Nowadays, the standard definition of such a business cycle is taken from the seminal work of Burns and Mitchell (1946). In particular, the cycle is defined as a combination of four periods: economic expansion, deceleration, recession/contraction and finally, recovery/acceleration. These cycles are certainly recurrent – with the prevailing recovery phase blending into the following expansionary period – but probably aperiodic, meaning that the exact length of each period will likely differ within each complete cycle. Figure 6.2 is a reproduction from the work of Van Vliet and Blitz (2011) illustrating the standard four-period economic or business cycle.

Business cycle identification and the application thereof in economics and finance continues to be a widely researched area. While most research uses a four-period cycle – with some decreasing this to two or three periods – there is far less con-

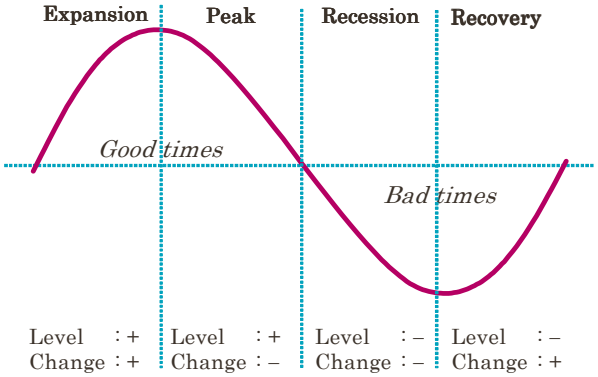


Figure 6.2. An economic cycle with four phases. Reproduced from van Vliet and Blitz (2011).

sistency in the economic variables chosen to identify these periods. This is to be expected though given the structural differences in the national economies. We thus turn towards the relevant South African literature to facilitate economic variable selection.

Moolman (2003) conducted one of the first studies of this kind using South African data, testing the ability of more than twenty economic indicators to predict turning points within the business cycle. Moolman found that short-term interest rates, the yield spread between 10-year and 3-month government bonds, and the composite index of leading indicators published by the South African Reserve Bank (SARB) were the best performing variables. Khomo and Aziakpono (2007) found similar results with regard to the predictive ability of the yield spread and showed that it had similar levels of predictive power to that of price momentum indicators (see Section 6.2.3 for more on this). A recent study by Mohapi and Botha (2013) also showed that the yield spread was able to accurately predict the 2008 sub-prime mortgage crisis in addition to all other major recessions dating back to 1980. It would thus appear that we have three prime candidates predicting the business cycle.

However, there is a problem here. The studies above all use some form of linear probit regression framework to consider whether lagged observations of the economic variables are significant predictors of an unobservable recession indicator variable. However, in order to answer this question, the studies all make use of the SARB's quarterly recession indicator variable. This means that the economic cycle has already been defined and the variables are being tested after the fact, which is somewhat putting the horse before the cart in terms of any ex ante regime-based applications.

Figure 6.3 compares the monthly drawdown series for South African equity against the regimes identified by a range of business cycle variables over the period January

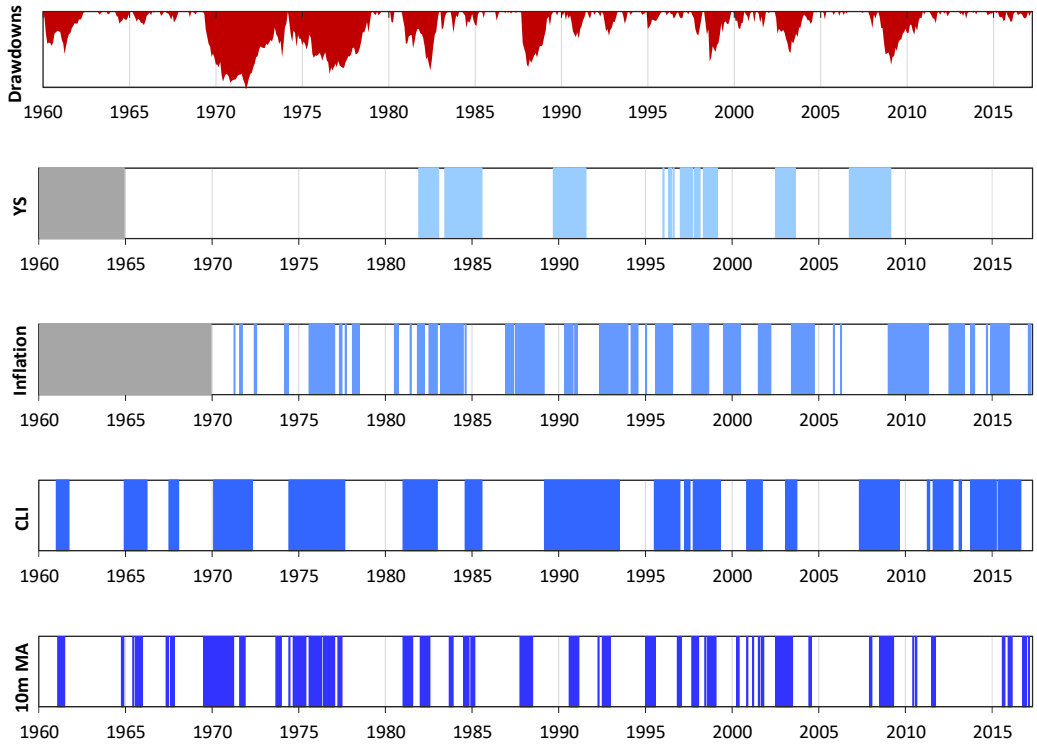


Figure 6.3. South African equity drawdowns versus business cycle recession profiles, Jan 1960 to Apr 2017.

1960 to April 2017. Motivated by our previous discussion, the economic variables include the 10-year to 3-month yield spread, 12-month changes in inflation and 12-month changes in the composite leading index (CLI). Periods are classified as recessions when the indicators take on negative values, displayed in the graphs as the respective shaded areas. We also include a final profile which defines recessions as those periods when the underlying equity index is below its 10-month moving average (MA). This technical price indicator is commonly used in current tactical asset allocation strategies and serves as a useful benchmark (Faber, 2013).

The regime profiles from the economic indicators are clearly quite different, both in terms of total frequency and average length. Although Mohapi and Botha (2013) showed that the yield spread accurately predicted all the SARB-indicated recessions back to 1980, if one extends the period back to 1965 (the start of available yield data), then a very different conclusion is reached regarding its predictive ability. The South African equity market was under water for the majority of the 1970s and early 1980s and there are two very large and obvious market crashes over this period. However, neither of these are flagged by the yield spread variable.

In comparison to the yield spread, the inflation indicator – which has data available back to 1970 – captures a portion of these early recessions prior to 1990 but still misses certain periods. We also observe that nearly half of the complete period is

flagged as recession, which is clearly at odds with the yield spread regimes and also with underlying economic rationale.

The CLI indicator also identifies a surprisingly high proportion of the full sample period as recession. Based on a graphical comparison with the equity drawdown curve, it would seem that the majority of the CLI-identified recessions occur close to the start of the actual market downturn but seem to continue quite far into the recovery phase. This is in contrast with the technical MA recessions, which line up with the CLI starting points but are considerably shorter and oscillate far more. A pertinent illustration of this divergence is the recessions – or lack thereof – identified from 1987 to 1994.

Table 6.1. In-sample recession and expansion statistics for South African equity returns.

	Recession Statistics				Expansion Statistics			
	Yield Spread	Inflation	CLI	MA	Yield Spread	Inflation	CLI	MA
Ave. Return	9.5%	15.8%	11.8%	-22.1%	20.5%	21.5%	24.2%	34.3%
Volatility	23.8%	22.7%	21.3%	24.5%	20.1%	19.8%	19.5%	16.4%
Sharpe Ratio	0.40	0.70	0.55	-0.90	1.02	1.08	1.24	2.09
Skewness	-0.50	-0.66	-0.50	-0.15	-0.39	-0.15	-0.37	0.11
Exc. Kurtosis	2.07	2.22	1.64	1.26	1.14	0.44	1.63	0.52
% No. Obs.	21.2%	44.3%	48.4%	28.4%	78.8%	55.7%	51.6%	71.6%

Regime-specific statistics for South African equity and bond returns calculated separately for each of the four indicators are given in Tables 6.1 and 6.2 respectively. Looking at the equity return statistics, the most important observation is that although the average returns in the recessionary periods are considerably lower than those seen in expansions, only the MA recession average is actually negative as one would expect. The difference in volatilities across the economic regimes are also not as extreme as that shown for the MA indicator. For the inflation and CLI regimes this is explained by once again considering the total number of observations identified within each regime. Around 44% and 48% respectively of all months are identified as recessions by these economic indicators. Assuming that there is in fact a dominant economic regime, it follows that the differences in these regime statistics are thus downwardly biased, albeit still directionally correct.

In comparison, the yield spread regime statistics are somewhat similar in value to those of the other two economic indicators, but only 21% of observations are now being classified as recessionary. Looking at Figure 6.3, we see that although the majority of these recession periods match up with the largest equity drawdowns, the inclusion of an incorrect recession from 1983 to 1986 and exclusion of the short but significant downturn in 1987 provides the downward bias.

Considering the bond returns in Table 6.2, it would appear that only the yield spread indicator accurately captures their regime-specific behaviour. This is to be expected given that the yield spread is intrinsically linked to bond performance and, furthermore, that negative yield spreads imply increasing bond prices. The inflation and MA indicators provide counter-intuitive results in that we see lower returns during recessions and higher returns during expansions, which is opposite to what economic theory suggests. This result is understandable in the case of the MA indicator though as this indicator is purely focussed on up- and down-trends in the underlying equity index. Finally, the CLI regimes provide economically reasonable results but only show minor differentiation when compared to the yield spread results.

Table 6.2. In-sample recession and expansion statistics for South African bond returns.

	Recession Statistics				**Expansion Statistics**			
	Yield Spread	Inflation	CLI	MA	Yield Spread	Inflation	CLI	MA
Ave. Return	15.7%	9.6%	11.6%	9.3%	9.4%	13.0%	9.3%	10.8%
Volatility	9.6%	7.6%	7.5%	7.5%	6.0%	7.1%	6.0%	6.4%
Sharpe Ratio	1.63	1.26	1.55	1.25	1.57	1.83	1.54	1.68
Skewness	-0.26	-0.71	-0.02	-0.37	-0.27	0.35	-0.29	0.12
Exc. Kurtosis	2.40	3.30	3.80	6.25	2.96	2.74	3.48	2.17
% No. Obs.	21.2%	44.3%	48.4%	28.4%	78.8%	55.7%	51.6%	71.6%

From this analysis, it would appear that macroeconomic indicators provide limited ability to partition equity markets relative to technical price indicators and that only the yield spread indicator is able to accurately capture statistically different regimes in the bond market.

6.2.2 Fundamental Equity Valuations

The second method for identifying market regimes is based on the underlying principle of value investing. Fundamental equity data is used to identify whether markets are currently undervalued or overvalued relative to historical norms. Undervalued markets should unlock value over time as they revert to their correct long-term equilibrium valuation, while overvalued markets are similarly likely to fall down to this level. An example of such an indicator is the cyclically adjusted price-to-earnings ratio (CAPE) introduced by Campbell and Shiller (1988). The CAPE ratio is commonly referenced in day-to-day market commentary as a measure of equity market health and extreme highs are taken as being indicative of imminent market crashes. Apart from the CAPE ratio, one can also use the full suite of value metrics – for example, price-to-book, dividend yield and price-to-cashflow – in a similar fashion and create regime identification rules based on the value spreads relative to some historical average.

As a simple test, we consider the equity regimes identified by a PE ratio and dividend yield combination on the FTSE/JSE All Share Index (ALSI) back to September 1986. The market is categorised into four states as given in Figure 6.4, based on whether the current PE ratio and dividend yield are above or below their respective historical running averages.

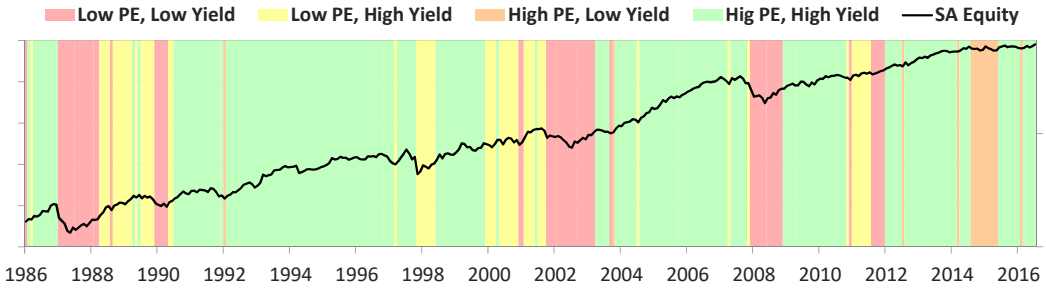


Figure 6.4. SA equity performance across fundamental valuation regimes, Sep 1986 to Apr 2017.

Based on the regime-specific statistics given in Table 6.3, there is a clear difference in realised volatilities across the high and low PE regimes and a similarly clear difference in average returns over the high and low yield regimes. The combination of the two metrics thus seems to provide a complete method for classifying over- and under-valued markets. It is worth bearing in mind though that only 4% of the total sample period falls within the *High PE, Low Yield* category, meaning that these estimated statistics are likely to be quite noisy.

Table 6.3. In-sample valuations-based regime statistics for South African equity returns, Sep 1986 to Apr 2017.

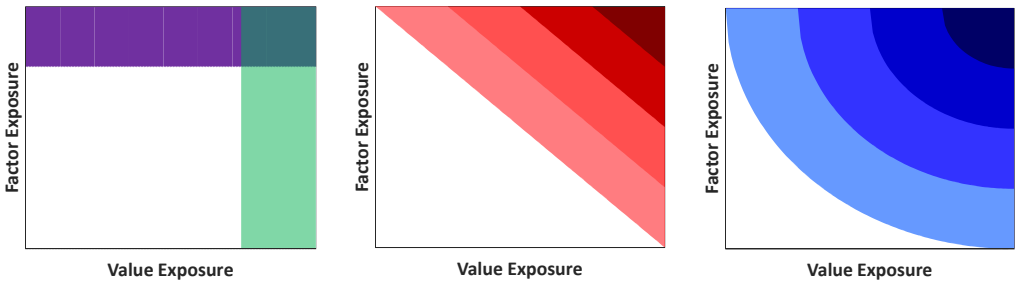
	Low PE, Low Yield	Low PE, High Yield	High PE, Low Yield	High PE, High Yield
Average Return	-11.72%	14.36%	-12.00%	22.11%
Volatility	24.46%	23.45%	11.68%	15.90%
% No. Obs.	17%	16%	4%	64%

Recently, the idea of relative valuation and the use of value spreads has received considerable attention in the equity factor space. This is largely due to many investors looking for ways in which to time factor exposures. Rob Arnott and his co-authors at Research Affiliates – one of the major investment firms in the Smart Beta space – produced a series of online white papers in 2016 on this issue, claiming that factor timing using relative valuation really does work. In response, Cliff Asness and the team from AQR Capital – another large Smart Beta firm – have written several pieces in which they categorically disagree with Arnott et al.'s (2016) findings and rather advocate holding a static diversified factor portfolio.

Our views on the use of value spreads to time factors are aligned more with Asness et al. (2017), for two reasons. Firstly, there is likely to be some degree of dependence

between the returns of a value-timed factor and the value factor itself. Investors may thus unknowingly increase their value exposure. Secondly and more importantly, using historical value spreads to time factors that are value-agnostic by construction equates to creating unintentional multi-factor portfolios. Such portfolios are likely to be sub-optimal relative to explicit multi-factor portfolios.

Timing factors based on their relative valuations is analogous to creating a multi-factor portfolio via the portfolio mixing method discussed in Section 5.5.1 of Chapter 5. This is depicted graphically in Figure 6.5a. However, rather than holding both factors in some fixed proportion as would be standard in portfolio mixing, the replicating multi-factor strategy for relative value timing would be to only hold those stocks within the overlapping exposure area in the top right corner of the panel. Periods when there are no stocks in this area then equate to 'sell' or 'under-weight' signals from the relative value indicator.



(a) Portfolio mixing **(b)** Signal averaging **(c)** Signal blending

Figure 6.5. Multi-factor portfolio construction methods.

Rather than unintentionally following this very specific and likely concentrated approach to multi-factor construction, investors should instead explicitly define their multi-factor construction method. Furthermore, they should consider the use of the more robust integrated scoring approaches depicted in Figures 6.5b and 6.5c. The portfolio mixing method, while transparent, does not account for interactive factor effects nor does it allow for variable factor signal decay speeds. Integrated scoring approaches do account for these issues, with the additive signal averaging and multiplicative signal blending approaches given here being examples of simple but robust multi-factor construction methods.

In summary, the classification of the underlying equity market into over- and under-valued regimes shows promise. However, using a similar relative valuation framework to time factors does not seem particularly beneficial given that one ultimately creates sub-optimal multi-factor portfolios. Investors should rather consider explicit multi-factor construction methods such as signal averaging and signal blending.

6.2.3 Technical Indicators: Momentum and Implied Volatility

Apart from considering the macroeconomic or fundamental classifications given above, one can also partition markets based on a range of technical or quantitative indicators. Such indicators are commonly used in systematic trading strategies as a means of tactically scaling risk exposure across a range of sectors or asset classes. Successful indicators are those which have strong theoretical and/or behavioural motivations and whose predictive power stems from taking advantage of a particular characteristic or stylised fact of the underlying return distribution.

Flint et al. (2014) tested the ability of a range of such indicators to accurately predict South African equity index regimes under the assumption of a two-state regime model, comprising a down-trending, volatile market and an up-trending, stable market. It was found that a number of these timing indicators accurately identified the major equity drawdowns since the mid-1990s and produced compelling results when utilised in a systematic timed hedging strategy. In particular, indicators based on probabilistic momentum and implied volatility were found to have the highest predictive and practical value.

Probabilistic momentum, introduced by Varadi (2014), translates monthly excess market returns over a specified historical period into a probability of outperforming cash (or another specified asset). If the probability of outperformance is lower than a given threshold k, then the market is said to be down-trending. If the outperformance probability is greater than 1–k then the market is said to be up-trending. Assuming that $k < 50\%$ there is therefore a 1–$2k$ buffer that needs to be crossed before a regime change is signalled. This buffer range ensures that weak or incorrect momentum signals are ignored.

Mathematically, the probability of outperformance, PM_t, is calculated by transforming excess monthly market returns in a given period into a t-score and using the Student's t distribution, ϕ, with n–1 degrees of freedom to convert this score into a probability,

$$PM_t = \phi\left(\frac{\mathbb{E}\left(R_m - r_f\right)}{\sqrt{\frac{\mathbb{V}ar\left(R_m - r_f\right)}{n}}}, n-1\right). \tag{6.1}$$

The similarities to the Sharpe Ratio are clear and thus probabilistic momentum can be thought of as a risk-adjusted momentum indicator. In the classifications below, we use rolling 8-month periods and a threshold value of $k = 30\%$. This means that if $PM_t < 30\%$, it will have to move above 70% before a regime change is recognised, and vice versa for $PM_t > 70\%$.

The implied volatility indicator is considerably simpler and is based on the stylised facts of volatility clustering and the inverse relationship between returns and volatility. These characteristics ensure that down-trending markets will coincide with high volatility and that this period should be somewhat persistent. The indicator is thus defined as whether 3-month at-the-money implied volatility is above or below a given historical percentile. Based on the work of Flint et al. (2014), we set this threshold to be the rolling historical 70^{th} percentile.

Figure 6.6 displays the combined four-state market classification as well as the time series of the individual indicators over the period February 1996 to April 2017. Despite significant overlap, the periods in which the indicators denote contradictory regime signals are quite noticeable. For example, the absence of a high volatility regime during the 2002/2003 period, when markets were strongly down-trending, is an important reminder that the manner in which a downturn occurs can have a strong effect on the predictive capacity of the underlying technical indicators.

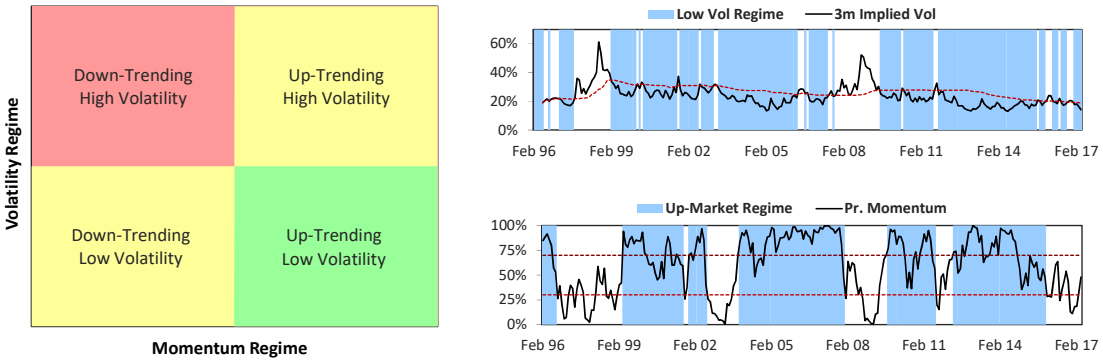


Figure 6.6. Technical indicator classification system and historical South African indicator profiles, Feb 1996 to Apr 2017.

Another observation from Figure 6.6 is that implied volatility moves considerably faster than probabilistic momentum, meaning that although momentum regimes will be more stable – and thus generate less turnover – this will come at the cost of missing the initial phases of any crash or recovery.

Table 6.4 gives the in-sample regime-specific statistics for the equity factor universe, and Table 6.5 gives similar regime statistics for the balanced asset class universe. The percentage in brackets next to each regime heading represents the number of months classified in that regime relative to the full sample. In line with Russo (2015), we report the p-values of a paired t-test of the return differences between each factor and the average returns of the remaining factors. For each regime, we show the annualised average return and volatility of the factor returns along with the Sharpe ratio, return skewness and excess kurtosis, and finally the minimum and maximum monthly returns.

Let us first discuss the factor statistics in Table 6.4. It is not surprising to see that the up-trending, low volatility regime shows high average factor returns and generally low realised factor volatilities across the entire universe, or that one sees negative average factor returns and elevated realised factor volatilities in the down-trending, high volatility regime. What is surprising though is that during up-trending, high volatility regimes, we see even stronger negative returns but at very low realised volatilities, while during down-trending, low volatility regimes we see the highest recorded average returns but coupled with fairly high realised volatilities.

To understand these results, let us contextualise the timing of the four technical-based regimes in terms of the stylised business cycle given in Figure 6.2. Probabilistic momentum will lag markets by construction, while implied volatility is arguably one of the best forward-looking estimates of market risk. Putting these two facts together, one is likely to see a positive momentum and high implied volatility combination during markets that have already peaked and are beginning to decelerate. And while uncertainty would be high during such a time, realised volatility may still be low as markets start to account for the possibility of a future recession. In such a situation one would also expect negative skewness to dominate the distribution, which is exactly what is given in the table.

In a similar vein, the combination of negative momentum and low implied volatility is likely to be evident at the start of any recovery period. This is when markets would be accelerating at their quickest rate – hence the strong positive returns in the table – but would also still experience a high degree of realised volatility as market participants phase out existing defensive holdings in favour of growth assets. Despite this, it is still quite remarkable to see just how strong the returns are during this initial part of the recovery phase.

In addition to these general trends highlighted above, there are also some factor-specific nuances within the various regimes. Such a classification makes it clearer to identify when each factor is rewarded or unrewarded. For instance, momentum shows the least negative return of -16.5% during the decelerating component of the economic cycle – i.e. up-trending but high volatility – but contrastingly displays the worst return of -23.5% during the true recessionary periods – i.e. down-trending and high volatility. In contrast, the low beta factor is the least affected during market crashes, showing a return of only -10.4%, but instead records the worst loss by a considerable margin of -36.4% during up-trending but high volatility market. Note, however, that the number of observations within this regime is considerably lower than in any of the other regimes and so there is bound to be some noise in the individual statistics.

Table 6.4. In-sample probabilistic momentum and implied volatility regime statistics for long-only SA equity factor returns, Jan 2003 to Apr 2017.

Up, Low Vol (62%)	Size	Value	Profitability	Investment	Momentum	Low Volatility	Low Beta
Average Return	26.4%	23.3%	29.8%	25.9%	34.3%	29.3%	29.0%
p-value	0.06	**0.00****	0.33	0.26	**0.00****	0.66	0.64
Volatility	12.0%	14.1%	12.3%	13.7%	13.9%	10.5%	11.7%
Sharpe Ratio	2.21	1.66	2.42	1.90	2.47	2.80	2.49
Skewness	0.17	0.03	-0.09	0.10	0.22	0.04	0.30
Exc. Kurtosis	-0.40	-0.05	0.25	-0.16	0.10	-0.31	1.56
Min. Return	-4.6%	-9.2%	-7.9%	-6.4%	-6.4%	-4.3%	-6.6%
Max. Return	11.6%	11.8%	10.9%	12.4%	14.4%	10.1%	14.7%

Up, High Vol (6%)	Size	Value	Profitability	Investment	Momentum	Low Volatility	Low Beta
Average Return	-21.8%	-17.6%	-30.2%	-23.2%	-16.5%	-28.8%	-36.4%
p-value	0.42	0.22	0.42	0.76	0.25	0.58	0.22
Volatility	8.7%	8.7%	12.2%	8.1%	9.1%	9.4%	10.6%
Sharpe Ratio	-2.49	-2.01	-2.47	-2.86	-1.80	-3.06	-3.44
Skewness	-0.55	-0.36	-0.97	-0.58	-0.45	0.11	-0.81
Exc. Kurtosis	-0.70	-0.42	1.00	-0.08	-1.26	-0.67	0.61
Min. Return	-5.9%	-5.9%	-9.9%	-6.5%	-5.0%	-6.8%	-9.0%
Max. Return	1.6%	2.2%	2.2%	1.0%	1.9%	1.9%	1.5%

Down, Low Vol (15%)	Size	Value	Profitability	Investment	Momentum	Low Volatility	Low Beta
Average Return	41.8%	39.5%	35.9%	41.5%	41.3%	26.7%	28.2%
p-value	0.08	0.29	0.96	0.21	0.21	0.06	0.09
Volatility	15.6%	16.1%	14.6%	17.0%	13.9%	11.5%	13.0%
Sharpe Ratio	2.68	2.45	2.46	2.44	2.96	2.31	2.17
Skewness	0.27	-0.23	0.53	0.30	-0.07	0.05	-0.57
Exc. Kurtosis	0.34	-0.17	0.02	-0.47	-0.32	0.83	1.44
Min. Return	-5.0%	-6.9%	-4.6%	-4.5%	-4.1%	-5.9%	-7.7%
Max. Return	13.7%	11.4%	12.1%	13.9%	11.9%	9.3%	9.8%

Down, High Vol (17%)	Size	Value	Profitability	Investment	Momentum	Low Volatility	Low Beta
Average Return	-13.5%	-13.8%	-12.2%	-15.2%	-23.5%	-14.8%	-10.4%
p-value	0.74	0.82	0.48	0.94	0.15	0.97	0.48
Volatility	19.8%	18.3%	19.2%	23.9%	20.4%	14.6%	16.6%
Sharpe Ratio	-0.68	-0.75	-0.63	-0.64	-1.15	-1.01	-0.63
Skewness	-0.06	-0.30	0.18	0.36	0.28	-0.90	-0.88
Exc. Kurtosis	-0.57	0.33	-0.41	0.38	0.08	1.48	1.12
Min. Return	-13.5%	-13.6%	-10.9%	-14.4%	-14.1%	-13.8%	-14.8%
Max. Return	9.5%	9.8%	11.2%	16.1%	12.4%	5.5%	7.2%

We now turn our attention to Table 6.5 and the balanced asset class regime statistics. Note that international asset class returns are calculated in dollar terms and that we include the exchange rate as a separate asset in order to isolate any regime-specific currency effects. The local currency comparative values can be approximated fairly accurately though by simply adding the asset asset and currency values together

As with the factor statistics, we see clear and significant differences in values across the four regimes, most of which are in line with our economic expectations. The asset classes generally perform best during the two low volatility regimes, and realised volatility only significantly increases for all asset classes in the down-trending, high volatility regime. Local and international bonds provide the best relative performance during the two high volatility regimes but otherwise underperform the riskier asset classes. The currency average return values tend to move in opposition to the rest of the asset classes across the regimes owing to the ratioed nature of the underlying exchange rate. Positive returns thus imply a weakening local currency.

We also note significant negative returns in the up-trending but high volatility regime and similarly significant positive returns during the down-trending but low volatility regime. Again, this is due to the relationship between these two regimes and the deceleration and recovery periods of the business cycle.

Perhaps most importantly, and in contrast to the results from the factors which are all equity-based portfolios, we see that at least two asset classes display significantly different average returns (p-value < 0.05) within each of the respective regimes. This suggests two things; firstly, that a strategy which tilts between asset classes on a regime basis should have the potential to considerably outperform a static multi-asset portfolio, and secondly, that the selected technical indicators do indeed classify markets into materially different and economically useful regimes.

6.2.4 Statistical Regime Models and Financial Turbulence

The final approach that we consider is statistical regime-switching models and, in particular, the hidden Markov model (HMM) introduced by Hamilton (1989). We consider a simple two-state regime-switching model of returns as an illustration. We can model the return r_t at time t in the following way:

$$r_t = \mu_{s_t} + \sigma_{s_t}\varepsilon_t, \quad \varepsilon \sim \mathcal{N}(0,1) \tag{6.2}$$

where μ_s and σ_s are the mean and volatility of the process, $s_t \in \{0,1\}$ is a binary state indicator, and ε_t is an *iid* random noise term. In this model, we assume that the two regimes are represented by two normal distributions and are thus fully specified

Table 6.5. In-sample probabilistic momentum and implied volatility regime statistics for the balanced universe asset class returns, Jan 2003 to Apr 2017.

Up, Low Vol (59%)	SA Equity	SA Bonds	SA Property	Int'l Equity	Int'l Bonds	Commodities	USDZAR
Average Return	26.6%	12.3%	28.5%	10.6%	3.7%	5.2%	6.0%
p-value	**0.00*****	0.67	**0.00*****	0.29	**0.00*****	0.06	0.09
Volatility	14.5%	6.9%	16.4%	11.7%	5.1%	12.8%	14.1%
Sharpe Ratio	1.83	1.78	1.74	0.9	0.74	0.4	0.42
Skewness	0.21	-0.11	-0.27	-0.21	-0.17	-0.33	0.74
Exc. Kurtosis	-0.12	0.85	0.99	0.06	0.17	0.54	1.36
Min. Return	-8.2%	-5.0%	-11.1%	-8.7%	-3.9%	-10.8%	-8.3%
Max. Return	13.1%	6.8%	15.8%	9.5%	3.7%	10.4%	16.8%

Up, High Vol (6%)	SA Equity	SA Bonds	SA Property	Int'l Equity	Int'l Bonds	Commodities	USDZAR
Average Return	-32.2%	2.1%	-26.8%	-34.2%	5.0%	-13.2%	34.7%
p-value	0.06	0.11	0.19	**0.01***	**0.03***	0.44	**0.00****
Volatility	14.8%	7.4%	14.5%	12.2%	4.5%	12.9%	13.20%
Sharpe Ratio	-2.18	0.29	-1.84	-2.79	1.11	-1.02	2.62
Skewness	-0.76	0.02	-1.66	-0.75	-0.63	-0.33	0.61
Exc. Kurtosis	1.26	0.34	3.62	-0.35	0.15	-0.01	0.43
Min. Return	-13.1%	-3.6%	-13.9%	-10.1%	-2.4%	-8.2%	-3.4%
Max. Return	3.4%	4.2%	3.2%	1.1%	2.2%	5.6%	11.5%

Down, Low Vol (16%)	SA Equity	SA Bonds	SA Property	Int'l Equity	Int'l Bonds	Commodities	USDZAR
Average Return	28.8%	18.1%	13.2%	25.40%	6.6%	9.3%	-8.8%
p-value	**0.01****	0.19	0.99	0.06	0.22	0.8	**0.01***
Volatility	15.0%	5.8%	14.1%	14.7%	5.6%	10.7%	15.7%
Sharpe Ratio	1.93	3.11	0.93	1.73	1.17	0.87	-0.56
Skewness	0.57	0.85	0.21	-0.73	0.28	0.23	0.55
Exc. Kurtosis	-0.08	2.42	1.69	1.4	1.27	0.14	1.26
Min. Return	-5.0%	-1.8%	-9.5%	-11.2%	-3.4%	-6.0%	-9.9%
Max. Return	14.1%	7.30%	12.9%	10.6%	4.9%	8.3%	12.6%

Down, High Vol (19%)	SA Equity	SA Bonds	SA Property	Int'l Equity	Int'l Bonds	Commodities	USDZAR
Average Return	-14.2%	9.4%	-7.1%	-11.8%	6.4%	-19.5%	16.0%
p-value	0.21	**0.02***	0.68	0.26	**0.05***	**0.04***	0.14
Volatility	26.3%	12.3%	24.2%	21.7%	7.5%	20.5%	20.4%
Sharpe Ratio	-0.54	0.77	-0.29	-0.54	0.86	-0.95	0.78
Skewness	-0.69	-0.24	0.91	-0.6	0.2	-0.43	0.59
Exc. Kurtosis	2.64	1.66	2.01	0.61	0.3	2.95	1.18
Min. Return	-29.3%	-10.5%	-14.5%	-19.6%	-3.8%	-22.3%	-11.9%
Max. Return	14.0%	9.2%	22.8%	11.5%	6.6%	13.8%	18.5%

by the mean and volatility parameters. The (unconditional) probability of being in regime $s_t = 0$ is given by π_0 and the probability of being in regime $s_t = 1$ is thus given by $1 - \pi_0$. Regime switches are governed by the transition probability matrix, $P_t = \{p_{ij,t}\}$, which represents the probability of transitioning from regime i at time $t - 1$ to regime j at time t. These probabilities can be fixed or time-varying depending on the model.

The mixture of normal distributions displayed in Figure 6.1 was produced by this model. It is appealing because it allows one to match the most important statistical properties of empirical returns – negative skew, fat tails and volatility clustering – in a fairly simple manner. The reason for this being called a hidden Markov model is because the actual regime variables are unobservable and need to be inferred from the available data. By fitting this model to empirical data, we thus attempt to estimate the underlying characteristics of each regime as well as the manner in which these regimes interact over the sample period.

Ang and Bekaert (2002) were one of the pioneers of regime-based asset allocation. Since then, there have been numerous studies making use of HMMs in portfolio and risk management applications – see Nystrup (2014) and Homescu (2015) for an overview. The majority of regime-switching models are applied directly to underlying asset or factor returns, but there are a few papers that apply the model to other variables such as statistical or economic indicators. In this work, we consider the latter approach and apply a regime-switching model to the financial turbulence index.

The financial turbulence index was proposed by Chow et al. (1999) and applied in portfolio optimisation and asset allocation by Kritzman et al. (2012). Chow et al. (1999) define a turbulent market as one in which assets behave in an 'unusual' fashion. Unusualness includes any period which shows either high volatility relative to the historical norm or different correlations relative to the historical norm, or a combination of both features. In order to capture all facets of unusualness, Chow et al. (1999) proposed the squared Mahalanobis distance as a measure of financial turbulence. Letting μ be the mean return vector of the asset universe and Σ its covariance matrix, we can therefore define turbulence d_t at time t as

$$d_t = (r_t - \mu)\,\Sigma^{-1}\,(r_t - \mu)^{'}. \tag{6.3}$$

Intuitively, turbulence can be thought of as the multivariate version of the univariate z-score. However, rather than dividing each asset's mean deviation by volatility alone, multivariate measures need to account for the correlations between the assets as well; hence the use of the covariance matrix as divisor.

One can then separate historical returns into quiet and turbulent regimes based on the turbulence index, either by using a fixed threshold level (Kritzman & Li, 2010) or through the use of a regime-switching model (Kritzman et al., 2012).

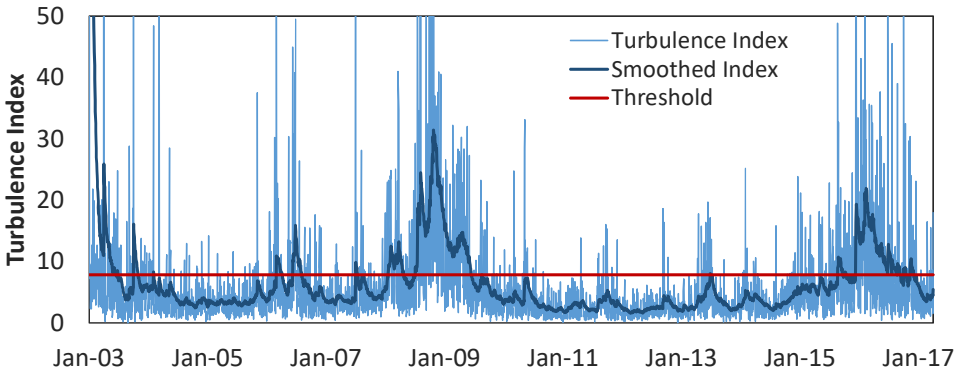


Figure 6.7. Factor universe turbulence index, Jan 2003 to Apr 2017.

As a demonstration, Figure 6.7 displays the daily turbulence index for the factor universe over the period January 2003 to March 2017. A smoothed index using an exponentially-weighted moving average is overlaid and we also include as a threshold level the 75^{th} percentile of a Chi-squared distribution with degrees of freedom equal to the number of factors. Quiet periods are defined as any periods when the smoothed index is below the threshold, while turbulent periods are those above the threshold.

Table 6.6 gives the average return, volatility and value-at-risk factor statistics computed from the identified quiet and turbulent regimes. From this, it is clear that this indicator creates regimes with statistics that are exactly in line with the intuitive definition of a turbulent market but also, and perhaps more importantly, in line with our economic intuition. Statistical regime-switching models based on the turbulence index thus appear to be compelling candidates for estimating and predicting market regimes.

Table 6.6. In-sample turbulence index regime statistics for SA equity factor returns, Jan 2003 to Apr 2017.

	Quiet Regime Statistics			Turbulent Regime Statistics		
Factor	Ave. Return	Volatility	Value-at-Risk	Ave. Return	Volatility	Value-at-Risk
Size	23.6%	10.0%	7.5%	-6.7%	24.4%	-37.4%
Value	21.8%	11.6%	2.8%	-8.8%	28.6%	-42.8%
Profitability	26.3%	12.0%	6.8%	-12.2%	28.6%	-44.7%
Investment	21.7%	12.2%	1.7%	-3.6%	30.9%	-41.9%
Momentum	29.5%	12.7%	9.0%	-13.9%	29.6%	-46.5%
Low Volatility	25.7%	9.9%	9.9%	-21.4%	24.2%	-45.8%

6.3 Tactical Allocation for Equity Factors and Balanced Portfolios

Although Section 6.2 showcases a number of methods and variables for successfully identifying market regimes, these results are mostly in-sample and do not shed much light on whether one can successfully incorporate regimes into an out-of-sample asset allocation framework. To answer this question, we conduct two practical tests. The first uses the technical probabilistic momentum and implied volatility indicators to tactically tilt towards assets that are expected to outperform and tilt away from underperforming assets. For this test, we make use of the equity factor universe and compare the performance of a tactical strategy against an equal-weight factor benchmark, which is rebalanced annually. We apply the regime-specific factor weight tilts given in Table 6.7 based on the technical indicator rules given in Section 6.2.3, with the only difference now being that the regime signal is used as an identifier for the *following* month.

Table 6.7. Factor weight tilts across technical indicator regimes.

Regime	Size	Value	Profitability	Investment	Momentum	Low Vol.	Low Beta
Up, Low Vol		-10%			+10%		
Up, High Vol			+5%		+5%	-5%	-5%
Down, Low Vol	+5%				+5%	-5%	-5%
Down, High Vol	+5%			-5%	-5%		+5%

Table 6.8 displays the performance results for the tactical and equal-weight factor portfolios. The tactical portfolio returns are on average 0.9% higher than the benchmark per annum, which is a highly significant improvement (p-value < 0.001). And while this differential does not explicitly account for trade fees, based on the actual two-way portfolio turnover, we calculate that the total transaction costs required to zero this differential are 1.61% per rebalance, which is extremely high. In addition to the return improvements, there is also a slight reduction in volatility, leading to improved risk-adjusted performance relative to the benchmark. In light of the extremely high correlations between the long-only factor returns over this historical period, the overall improvements from the tactical factor allocation framework therefore appear to be quite meaningful.

For the second test, we calculate turbulence based on the return series from the balanced universe and fit a regime-switching model to a growing data window in order to predict the upcoming regime and thus tilt the balanced portfolio accordingly. This tactically tilting portfolio is compared to a strategic asset allocation of 40% local equity, 10% local property, 10% foreign equity, 35% local bonds and 5% foreign bonds, which is rebalanced annually.

Table 6.8. Equal-weight versus tactical allocation strategy performance with the factor universe, Jan 2003 to Mar 2017.

	Equal-Weight Portfolio	TAA Portfolio
Average Return	18.69%	19.59%
p-value	n.a.	0.001***
Volatility	14.01%	13.89%
Return-Risk Ratio	1.3	1.41
Skewness	-0.19	-0.14
5% VaR (monthly)	-4.65%	-4.75%
Max. Drawdown	-31.74%	-29.76%
Ave. 2-way Turnover pa	5.2%	32.7%
Break-Even Cost per Rebalance	n.a.	1.61%

The regime-based weight tilts in Table 6.9 for the balanced universe are considerably larger than those for the factor universe because of the far larger differences in the performance of the various asset classes over the two regimes. Performance results for the strategic and tactical strategies are given in Table 6.10. Note that we conduct this test from the viewpoint of a local investor and thus convert the foreign assets into local currency.

Table 6.9. Factor weight tilts across technical indicator regimes.

Regime	SA Equity	SA Bonds	SA Property	Global Equity	Global Bonds
Quiet	+10%	-10%	+5%		-5%%
Turbulent	-10%	+15%		-10%	+5%

We again see that the regime-based tactical allocation provides a significant improvement in average return of 2.4% relative to the strategic benchmark and that the estimated break-even transaction costs per rebalance are extremely high at 2.71%. While the improvement in return does come at the cost of a slight increase in volatility, we still find that the balanced tactical asset allocation (TAA) strategy meaningfully outperforms its benchmark on a risk-adjusted basis.

Based on the results of the tactical strategies presented above – which are only simple illustrations of the underlying thesis – we can thus conclude that a regime-based framework is definitely capable of adding value to an asset allocation process.

Table 6.10. Strategic versus tactical allocation strategy performance with the balanced universe, Jun 1995 to Apr 2017.

	Equal-Weight Portfolio	TAA Portfolio
Average Return	12.27%	14.65%
p-value	n.a.	0.000***
Volatility	11.89%	12.00%
Return-Risk Ratio	1.03	1.22
Skewness	-0.88	-0.38
5% VaR (monthly)	-4.75%	-4.08%
Max. Drawdown	-30.41%	-26.20%
Ave. 2-way Turnover pa	10.4%	71.7%
Break-Even Cost per Rebalance	n.a.	2.71%

6.4 Conclusion

The use of regimes and regime-switching models is becoming increasingly popular in finance. The reasons for this are because regime-based frameworks align with the observed nonlinear and non-normal market dynamics, are intuitive in their underlying economic narrative, and provide significant modelling power in a simple and tractable manner. In this chapter, we considered four alternative methods for identifying market regimes; namely through macroeconomic variables, fundamental valuation metrics, technical market indicators and statistical regime-switching models. These methods were tested using a long-only equity factor universe and a representative balanced portfolio universe.

We found that the tested macroeconomic variables showed limited ability to partition equity markets but that the yield spread indicator was able to provide economically sensible bond return regimes. In contrast, we found that the valuation metrics that were tested were able to successfully partition the equity market into four different regimes. However, using relative valuations to similarly partition equity factors is analogous to creating some form of implicit multi-factor portfolio. Investors should rather consider explicit methods such as signal averaging or signal blending to create multi-factor portfolios if that is their goal.

Simple technical indicators based on probabilistic momentum and implied volatility were found to accurately partition both the factor and balanced data universes into up-trending, low volatility states and down-trending, high volatility states. We also showed that a simple two-regime switching model based on the turbulence index was able to accurately capture differing regimes in the underlying equity factor universe.

Based on these findings, we tested two out-of-sample regime-based asset allocation frameworks. The first test used technical indicator regimes to tactically tilt equity factor exposures, while the second used a regime-switching turbulence model to tactically tilt asset class exposures for a balanced portfolio. In both tests, we found that the regime-based tactical allocation strategies outperformed their static benchmarks on an absolute return and risk-adjusted return basis, suggesting that a regime-based framework can add value to the asset allocation process.

There exist several avenues for further research based on the work in this chapter, of which we will highlight two. Firstly, it would be interestesing to apply a regime-switching model directly to the underlying asset or factor returns and compare the performance of tactical strategies based on such an implementation versus the strategies given above. Secondly, the tactical strategies outlined above are, by our own admission, fairly simple representations of what is possible when using a regime-based framework. One could therefore extend this work and consider the performance of a suite of more complex tactical and dynamic asset allocation strategies based on regime-switching models, particularly in a South African investment setting.

7

Extending Risk Budgeting for Market Regimes and Quantile Factor Models

> *When you reach the end of the road, there's only one thing to do - build more road.*
>
> **— Ashleigh Brilliant**
> (Author and cartoonist)

Chapter Synopsis

In this chapter, we combine several disparate avenues in the literature to create a novel, unified risk-based optimisation framework. Specifically, we extend the existing risk budgeting approach of Richard and Roncalli (2015) to allow for changing market regimes, factor dependence and nonlinear and asymmetric market structure.

We show that the existing framework can be readily extended to include a factor-dependent return process using standard models available in the literature. Structural changes in market conditions are then incorporated into the framework through the use of a regime-switching turbulence index and the nonlinear and asymmetric market dependence structure is accounted for by using quantile factor models. Most importantly, this extended framework is only comprised of a series of linear models, and is thus simple to understand and implement. We consider two applications of the extended framework, namely scenario analysis and parameter uncertainty analysis, through means of a simple empirical case study.

Finally, we introduce the concept of Risk Maps, which provide managers with a graphical approach for estimating and evaluating risk optimality in a multi-objective and multi-scenario setting.

This chapter is adapted from the journal article by Flint and du Plooy (2018) and addresses research questions 2d, 4c and 4d as outlined in Section 1.2.3.

7.1 Introduction

The theory and practice of optimal portfolio construction is a primary concern in modern financial research. Nearly seventy years ago, Markowitz (1952) introduced the efficient frontier approach to asset allocation, and his is still the most popular framework for constructing portfolios of assets. Under this framework, an optimal portfolio is defined as the combination of assets that maximises the expected return of the portfolio for a specified level of portfolio risk at a given time horizon. In theory, then, the portfolio construction problem has been solved. One simply needs to input the asset returns' means and covariances into the framework to obtain an optimal portfolio specific to one's risk preferences. When applied in practice, however, the model is found to be very sensitive to small changes in the estimated means, resulting in the optimisation procedure outputting unreasonable or impractical allocations in almost all cases. This behavior led to Michaud (1989) coining the infamous phrase "error maximiser".

To address these weaknesses in the framework, academics and practitioners alike have focused their efforts on two main areas. The first area is based on creating better expected return estimates and has led to the rise of the factor investing paradigm; see Ang (2014) and Cochrane (2011), among others, for comprehensive reviews of this literature. The second area is based on all things risk related; namely, risk-based portfolio construction, improving the efficiency of the risk estimation, and creating new risk and diversification measures. In the last decade in particular, a number of significant advancements have been made in the risk budgeting space, and so risk- and diversification-based investment portfolios have become a common feature of markets worldwide.

In this research, we combine several disparate avenues in the literature to create a novel, unified risk-based optimisation framework. Specifically, we extend the existing risk budgeting approach of Richard and Roncalli (2015) to allow for changing market regimes, factor dependence and a nonlinear, asymmetric market structure. Market regimes are incorporated into the framework via the regime-switching turbulence index of Kritzman et al. (2012) and factor dependence is addressed with the use of a generalised factor risk estimation framework (Meucci, 2019). To account for the nonlinear, asymmetric market dependence structure, we also incorporate quantile regression into the risk modeling component (Chen et al., 2016). We describe the framework mathematically, detail its practical implementation and then demonstrate it using real-world market data. As an extension of this framework, we introduce the idea of Risk Maps, which provide managers with a graphical method for evaluating portfolio risk and diversification across multiple risk dimensions and market settings

simultaneously. Furthermore, Risk Maps also naturally provide one with a means of measuring portfolio risk optimality in a multi-objective, multi-scenario setting.

While risk-based portfolios have been extensively researched, to the author's best knowledge, the impact of regimes has not yet been systematically incorporated into the estimation methodology of these portfolios. Similarly, quantile regression has never been used in the construction of risk-based portfolios. We therefore contribute to the risk-budgeting literature by proposing a general framework that can account for the stylised facts of financial markets. This framework will ultimately enable practitioners to evaluate or build risk-based portfolios that are robust to changes in market structure, estimation error and nonlinear dependencies, while remaining analytically simple and easy to implement.

The rest of this chapter is organised as follows. Section 7.2 extends the risk budgeting framework mathematically for market regimes and quantile factor risk models, and describes a process for practically implementing this extended framework. Section 7.3 showcases two practical applications of the framework by means of a simple case study on a South African stock universe. Section 7.4 introduces the concept of Risk Maps and demonstrates their use with South African global minimum variance stock portfolios. Section 7.5 concludes and outlines some ideas for further research.

7.2 Extending the Risk Budgeting Framework

Let us start with the generalised risk budgeting framework of Richard and Roncalli (2015), who present the following constrained minimum variance optimisation problem:

$$\hat{w}\left(\lambda,\gamma,\delta,\kappa\right) = \operatorname{argmin}\left[\frac{1}{2}w\Sigma w' - \lambda\mathcal{D}\left(w,\gamma\right) + \left(\lambda+1\right)\mathcal{B}\left(w,\delta,\kappa\right)\right] \tag{7.1}$$

$$\begin{aligned} s.t. \quad & w \geq 0, \\ & \mathcal{D}\left(w,\gamma\right) = \gamma\sum_{i=1}^{n}\ln w_i - \left(1-\gamma\right)\sum_{i=1}^{n}w_i^2, \\ & \mathcal{B}\left(w,\delta,\kappa\right) = \sum_{i=1}^{n}w_i\left(\delta + \kappa\left(\Sigma w_{\text{cw}}\right)_i + \left(1-\delta-\kappa\right)\sigma_i\right), \\ & \left\{\gamma,\delta,\gamma+\kappa\right\} \in \left[0,1\right],\ \lambda \geq 0, \end{aligned}$$

where $w = \{w_i\}$ is the portfolio holdings vector, $\Sigma = \{\sigma_{ij}\}$ is the covariance matrix of the underlying portfolio instruments, $\mathcal{D}$ is a diversification constraint function and $\mathcal{B}$ is a budget constraint function. The parameter λ controls the impact of portfolio diversification, γ controls the trade-off between weight diversification and risk diversification, δ controls the budget allocation, and κ controls the impact

of any tracking error constraint. Taken together, the diversification and budget functions $\{\mathcal{D}, \mathcal{B}\}$ and the parameter set $\{\lambda, \gamma, \delta, \kappa\}$ ensure that the constrained optimisation problem above is flexible enough to encompass a wide range of the specific risk-based portfolios currently in vogue, treating them as special cases. In particular, these include the minimum variance (MV), equal weight (EW), equal risk contribution (ERC), most diversified portfolio (MDP), risk parity (RP), beta parity (BP) and capitalisation weight (CW) portfolios. Table 7.1 gives the parameter combinations required to produce these specific risk- and diversification-optimal portfolios. Furthermore, Richard and Roncalli (2015) show that by varying the parameters values between those given in Table 7.1, their generalised risk budgeting framework can also be used to create 'frontiers' of optimal portfolios, linking a number of the specific risk-based portfolios together.

Table 7.1. Parameter combinations required for special case risk-based portfolios, as per Richard and Roncalli (2015).

Parameters	MV	EW	ERC	MDP	RP	BP	CW
λ	0	∞	1	0	∞	∞	0
γ		0/1	1		1	1	1
δ	1	1		0	1	0	
κ	0	0		0	0	1	1

Note that the framework is not prescriptive in terms of portfolio instrument type and is therefore equally well suited to use with underlying single assets, subgroups of assets within the portfolio (e.g. equity sectors), or risk factors that are external to the portfolio. However, this is not to say that a risk-optimal portfolio calculated in terms of underlying assets will be equivalent to a similarly risk-optimal portfolio calculated in terms of extrinsic risk factors. This point is an important one and has been well made by Roncalli and Weisang (2016), amongst others. As a result, one is forced to specify not only the desired risk- or diversification-optimality criterion but also the underlying asset universe in which the portfolio should be measured, managed and ultimately optimised.

In its given form, the optimisation programme 7.1 caters for a range of risk and diversification metrics but is largely silent in terms of selecting a type of asset universe. While this decision will generally depend on a portfolio manager's mandate or management preference, one can still make explicit the connection between underlying assets and risk factors by using a standard factor modelling framework (Meucci, 2019). In particular, the return on an asset R_i can be modelled as a linear combination of a set of market risk factors:

$$R_i = \alpha_i + \sum_{j=1}^{m} \beta_i^j \mathcal{F}_j + \varepsilon_i, \tag{7.2}$$

where α_i is a constant, β_i^j is the sensitivity of asset i to factor j, $\mathcal{F}_j$ is the return on factor j, and ε_i is a zero-mean iid error term that is also independent from any of the risk factors. Using standard results from linear factor models and changing to matrix notation, for a market comprising n assets and m risk factors, we obtain

$$\mathbb{E}[R] = \mu = \alpha + \beta\mathbb{E}[\mathcal{F}] \tag{7.3}$$

$$\mathbb{C}ov[R] = \Sigma = \beta\Omega\beta' + D. \tag{7.4}$$

In Equations 7.3 and 7.4, μ represents the n-vector of expected asset returns, α is an n-vector of constants, β is the $n \times m$ factor loadings matrix, Σ the $n \times n$ asset covariance matrix, Ω the $m \times m$ factor covariance matrix, and D the diagonal $n \times n$ covariance matrix of the error terms, denoting the idiosyncratic risk contribution. Substituting Equation 7.4 into Equation 7.1 thus enriches the original risk budgeting framework by adding an explicit dependence to a set of m systematic factors.[1] This enrichment does come at a cost though as both the factor covariance and factor loadings matrices are generally unobservable and thus need to be estimated, usually from historical data. We now focus our attention on this process.

7.2.1 Incorporating Market Regimes

Efficient estimation of covariance matrices has been a focus of risk literature for quite some time. In particular, enhancements stemming from Bayesian shrinkage (Ledoit & Wolf, 2004), random matrix theory (Laloux et al., 2000) and advanced time-series modelling (Engle, 2002) have now become fairly standard means of dealing with estimation error and sampling issues. However, another branch of uncertainty that is less explored (though arguably more influential in terms of its effect on risk estimates) is that associated with structural changes in the underlying market conditions. While the enhancements highlighted above could potentially be used to address this issue, we consider an alternative framework that is perhaps better suited for this problem; namely, regime-switching models.

Ang and Timmermann (2012) identify three reasons for the popularity of a regime-based framework. Firstly, regimes are intuitive and can naturally fit the ex-post narratives that investors use to explain market moves. Estimated regimes are often found to tie up with low and high volatility periods, up- and down-trending return periods and changes in underlying macroeconomic policy. Secondly, regime models are capable of accurately capturing the nonlinear and non-normal stylised facts prevalent in markets worldwide (Cont, 2001). Consider a simple mixture of two normal distributions which can display significant negative skew as well as fat tails.

[1] As a special case, one can remove this dependence by simply specifying the risk factors as the underlying assets themselves. In this instance, β and D become identity and zero matrices respectively, and thus $\Sigma = \Omega$.

Thirdly, because regime models are generally constructed as a linear combination of (log)normal distributions, they are simple to understand and implement. A regime-based framework thus allows one to capture the intricacies of the market while still affording analytical tractability and familiarity.

The underlying premise of regime models is that the actual future market conditions will constitute some unknown blend of known market state conditions. From this, one can then create a comprehensive range of future market possibilities and thus gain insight into the entire distribution of future market conditions, rather than simply a single expectation. Ang and Bekaert (2002) pioneered the use of regime-based models in the asset allocation process. Since then, a number of studies have made use of regimes in portfolio and risk management applications.[2] The majority of regime-switching models are applied directly to underlying asset or factor returns but there are a few papers that apply the model to other variables such as statistical or economic indicators. In this work, we follow the latter approach, focusing on the financial turbulence measure proposed by Chow et al. (1999) and applied by Kritzman et al. (2012) in a portfolio optimisation setting.

Chow et al. (1999) define a turbulent market as one in which assets behave in an 'unusual' fashion. Unusualness here includes any period that shows

- high volatility relative to the historical norm,
- different correlations relative to the historical norm, or
- a combination of both these features.

In order to properly capture this unusualness, Chow et al. (1999) propose the squared Mahalanobis distance as a measure of financial turbulence. Using the notation as above, the turbulence index d_t at time t is thus calculated as

$$d_t = (R_t - \mu_t)\,\Sigma_t^{-1}\,(R_t - \mu_t)^T \tag{7.5}$$

Kritzman et al. (2012) use a simple two-state hidden Markov model (HMM) to model this turbulence index:

$$d_t = \mu_{s_t} + \sigma_{s_t}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1)\,, \tag{7.6}$$

where μ and σ are the regime-switching mean and volatility of the process, $s_t \in \{Q, T\}$ is a binary state indicator representing quiet and turbulent regimes respectively, and ε_t is an iid random noise term. The regimes are modelled by two normal

[2] Please see Nystrup (2014) and Homescu (2015) for an overview of this research.

distributions and are thus fully specified by the mean and volatility parameters. The (unconditional) probability of being in regime $s_t = Q$ is given by π_0 and the probability of being in regime $s_t = T$ is thus given by $1 - \pi_0$. Regime switches are governed by the transition probability matrix $P_t = \{p_{ij,t}\}$, which represents the probability of transitioning from regime i at time $t-1$ to regime j at time t. Maximum-likelihood estimates for the parameters in Equation 7.6 can be obtained from an expectation-maximisation algorithm, and the most likely sequence of states $S = \{s_t\}$ can then be determined from a decoding algorithm in conjunction with these estimated parameters. Typical choices for these are the Baum-Welch and Viterbi algorithms respectively (Zucchini et al., 2016).

Once the market state sequence vector has been obtained, regimes are then easily incorporated into the return generating factor model, given in matrix form as

$$R_s = \alpha_s + \beta_s \mathcal{F}_s + \varepsilon_s \tag{7.7}$$

$$\Sigma_s = \beta_s \Omega_s \beta_s' + D_s, \tag{7.8}$$

where the subscript $s = \{Q, T\}$ represents the market regime. Note that both the factor covariance and factor loadings matrices are assumed to be regime-dependent and the set of factors within each regime $\mathcal{F}_s$ need not be equivalent across regimes. Furthermore, covariance estimation enhancements (e.g. shrinkage) can still be applied on a regime-specific basis.

Substituting equation 7.8 into the optimisation problem 7.1 leads to portfolios that are risk-optimal for either quiet regimes or turbulent regimes. However, it is unlikely that one will know for certain which of these two regimes will prevail over the future period under consideration. Instead, one can obtain a range of possible future scenarios by blending the risk estimates from the two given market regimes. Chow et al. (1999) suggest an elegant procedure for producing a blended estimate that accounts for one's view on the likelihood of each regime occurring as well as one's relative risk aversion toward each regime. Focussing only on the risk aspect, the blended covariance matrix Σ_B is calculated as

$$\Sigma_B\left(p, \eta\right) = \left(1 - p\right) \eta_Q \Sigma_Q + p \eta_T \Sigma_T, \tag{7.9}$$

where Σ_Q and Σ_T represent the quiet and turbulent covariance matrices respectively, p is the probability of a turbulent regime occurring, and $\eta = \{\eta_Q, \eta_T\}$ are the normalised risk aversions towards each regime. The special cases of the quiet and turbulent covariance estimates are recovered when the regime risk aversion parameters are equivalent and p equals either 0 or 1.

7.2.2 Incorporating Quantile Factor Modelling

The introduction of the regime-switching framework adds a significant degree of flexibility to the overall return modelling process. However, Equation 7.7 ultimately remains a linear conditional mean model of R_s with respect to the factor set $\mathcal{F}_s$. It is now widely accepted that the dependence structure within financial markets is generally nonlinear and asymmetric (Embrechts et al., 2001; Ang & Chen, 2002). What this means is that the relationship between asset returns and risk factors within each regime may not actually be linear. If so, one then needs to generalise Equation 7.7 to allow for the possibility of heterogeneous factor effects across the modelled return distribution. To this end, we consider the use of a quantile factor modelling process (Chen et al., 2016).

Koenker and Bassett (1978) introduced quantile regression as a robust alternative to the classical ordinary least squares (OLS) regression. Rather than only considering the conditional mean, quantile regression separately models each conditional quantile of the asset return distribution as a linear function of underlying risk factors. Perhaps due to the prevalence of linear factor models, quantile regression has been applied to a wide range of financial estimation problems. The most obvious of these is the problem of estimating robust stock betas in the presence of heteroscedasticity and non-normality (Atkins & Ng, 2014). Related portfolio management applications include the use of quantile regression in return forecasting and portfolio construction (Ma & Pohlman, 2008) as well as portfolio attribution via returns-based style analysis (Bassett & Chen, 2001).

On the risk management side, quantile regression has found a natural home in the value-at-risk (VaR) literature, a field that is centered on modeling a specific distribution quantile (Engle & Manganelli, 2004). The technique has also been used to estimate nonlinear financial dependence structures (Sim, 2016) and measure contagion risk (Park et al., 2015). Despite the considerable breadth of its application, to the best of our knowledge, quantile regression has not yet been considered in risk budgeting in any form. Furthermore, regime-switching quantile regression models were only recently introduced by Ye et al. (2016), with their focus being on measuring global contagion risk. In this work, we incorporate quantile regression into our risk budgeting framework through the introduction of quantile-specific covariance matrices.

Let $F\left(R_i|\mathcal{F}\right)$ be the conditional cumulative distribution function of R_i given the factor set $\mathcal{F}$, and denote the τ^{th} conditional quantile as $\mathcal{Q}\left(\tau|\mathcal{F}\right) = \inf\left\{r : F\left(r_i\right) \geq \tau|\mathcal{F}\right\}$. The single-asset quantile factor model can then be written as

$$\mathcal{Q}\left(\tau|\mathcal{F}\right) = \alpha_i^\tau + \sum_{j(\tau)} \beta_i^{j,\tau} \mathcal{F}_j + \mathcal{Q}_\varepsilon\left(\tau|\mathcal{F}\right), \tag{7.10}$$

where $\mathcal{Q}_\varepsilon\left(\tau|\mathcal{F}\right)$ is the conditional τ-quantile of the error distribution, and the quantile-specific regression parameters are estimated by numerical optimisation of an absolute tilted loss function. The generalised quantile factor model also allows for quantile-specific factor selection, with Chen et al. (2016) providing a simple procedure for selecting factors and estimating loadings simultaneously.

Following from Section 7.2.1, we can incorporate regimes into the quantile factor model in Equation 7.10. Switching to matrix form for notational simplicity, we obtain

$$\mathcal{Q}_s^\tau = \alpha_s^\tau + \beta_s^\tau \mathcal{F}_s^\tau + \mathcal{Q}_{\varepsilon,s}^\tau, \tag{7.11}$$

where the subscript $s = \{Q, T\}$ represents the market regime as before, and $\mathcal{F}_s^\tau$ represents the regime- and quantile-specific factor set.

In the majority of cases where the factor set is the same across both regimes, quantile dependence is limited to the changes in the quantile-specific β_s^τ factor loadings. However, it is important to realise that the estimated conditional error quantile will also change in line with these regression parameters. Thus, if one uses the standard linear factor covariance representation as per Equation 7.4, then each quantile-specific covariance estimate in each regime will be exactly the same. This is because any changes in the systematic factor risk term, $\beta_s^\tau \Omega_s^\tau \left(\beta_s^\tau\right)'$, will be perfectly offset by similar but opposite changes in the idiosyncratic error risk term, $D^{(\tau)}$. Or to put it differently, while the ratio between factor and idiosyncratic covariance contributions will change across quantiles, the total covariance matrix will not. Therefore, in order to create covariance matrices that actually do differ across quantiles, we fix the idiosyncratic risk matrix at the median value $D_s^* = D_s^{(0.5)}$ of the relevant regime and estimate the quantile-specific covariance matrix as

$$\Sigma_s\left(\tau\right) = \beta_s^\tau \Omega_s^\tau \left(\beta_s^\tau\right)' + D_s^*. \tag{7.12}$$

The use of D_s^* means that we recover the original regime-specific covariance matrix when considering the median quantile. As with Equation 7.8, one can once again use enhanced covariance estimation techniques to calculate Ω_s^τ if desired.

Substituting Equation 7.12 into Equation 7.9, we finally obtain

$$\Sigma_B\left(p, \eta, \tau_Q, \tau_T\right) = \left(1-p\right)\eta_Q \Sigma_Q\left(\tau_Q\right) + p\eta_T \Sigma_T\left(\tau_T\right), \tag{7.13}$$

where the regime-specific quantile parameters τ_Q and τ_T, do not necessarily have to be equivalent. Relaxing the need for regime-specific quantile equality allows one to indirectly account for potentially nonlinear switches between regimes, while still making use of a purely linear blending procedure. Equation 7.13 can then be substituted into the risk budgeting optimisation 7.1 in order to find the optimal portfolio for a given risk metric, regime probability and conditional return quantile:

$$\hat{w}\left(\lambda, \gamma, \delta, \kappa; p, \eta, \tau_Q, \tau_T\right) = \operatorname{argmin}\left[\frac{1}{2} w \Sigma_B\left(p, \eta, \tau_Q, \tau_T\right) w' - \lambda \mathcal{D}\left(w, \gamma\right) + \left(\lambda + 1\right) \mathcal{B}\left(w, \delta, \kappa\right)\right] \tag{7.14}$$

$$s.t. \qquad w \geq 0.$$

7.2.3 Practical Implementation of the Extended Framework

Sections 7.2.1 and 7.2.2 mathematically describe how to extend the generalised risk budgeting framework for market regimes and quantile factor models. In this section we describe the process of implementing the framework, which practically means finding the covariance matrix of interest to plug into the risk budgeting optimisation. Let us start by considering the base case of estimating an OLS linear factor model from all available historical data and calculating a single covariance matrix, labelled Σ_0. This scenario is depicted in panel A of Figure 7.1.

Panel B then shows how one might practically add a regime-switching component to the covariance estimation procedure. Firstly, one would need to calculate the turbulence index over the full sample period and then fit the regime-switching model to this series. Once fitted, the time series of most likely states can then be used to partition the historical data sample into separate quiet and turbulent regimes. Once the regimes have been identified and partitioned, the regime-specific covariance matrices can be estimated by fitting separate OLS factor models to each regime. Finally, the input covariance matrix is calculated from a simple linear blending function incorporating the two regime-specific matrices. For the case where the factor set and risk aversion parameters remain the same for both regimes, the base covariance matrix Σ_0 is retrieved by setting the turbulent regime probability equal to the proportion of the data sample classified as turbulent.

In panel C, we show how one might practically incorporate quantile factor models into the covariance estimation procedure when a regime-switching framework has not been used. Firstly, one implements quantile regression using the full data sample to estimate a set of factor loadings matrices across a specified range of quantiles. The idiosyncratic risk matrix D^* is then fixed as equal to the median idiosyncratic risk matrix $D^{(0.5)}$. Quantile-specific covariance matrices are then calculated from

the quantile-specific factor loadings and factor covariance matrix as well as the fixed idiosyncratic risk matrix. When the factor set remains constant across the quantiles, the base covariance matrix Σ_0 is equivalent to the median quantile covariance matrix.

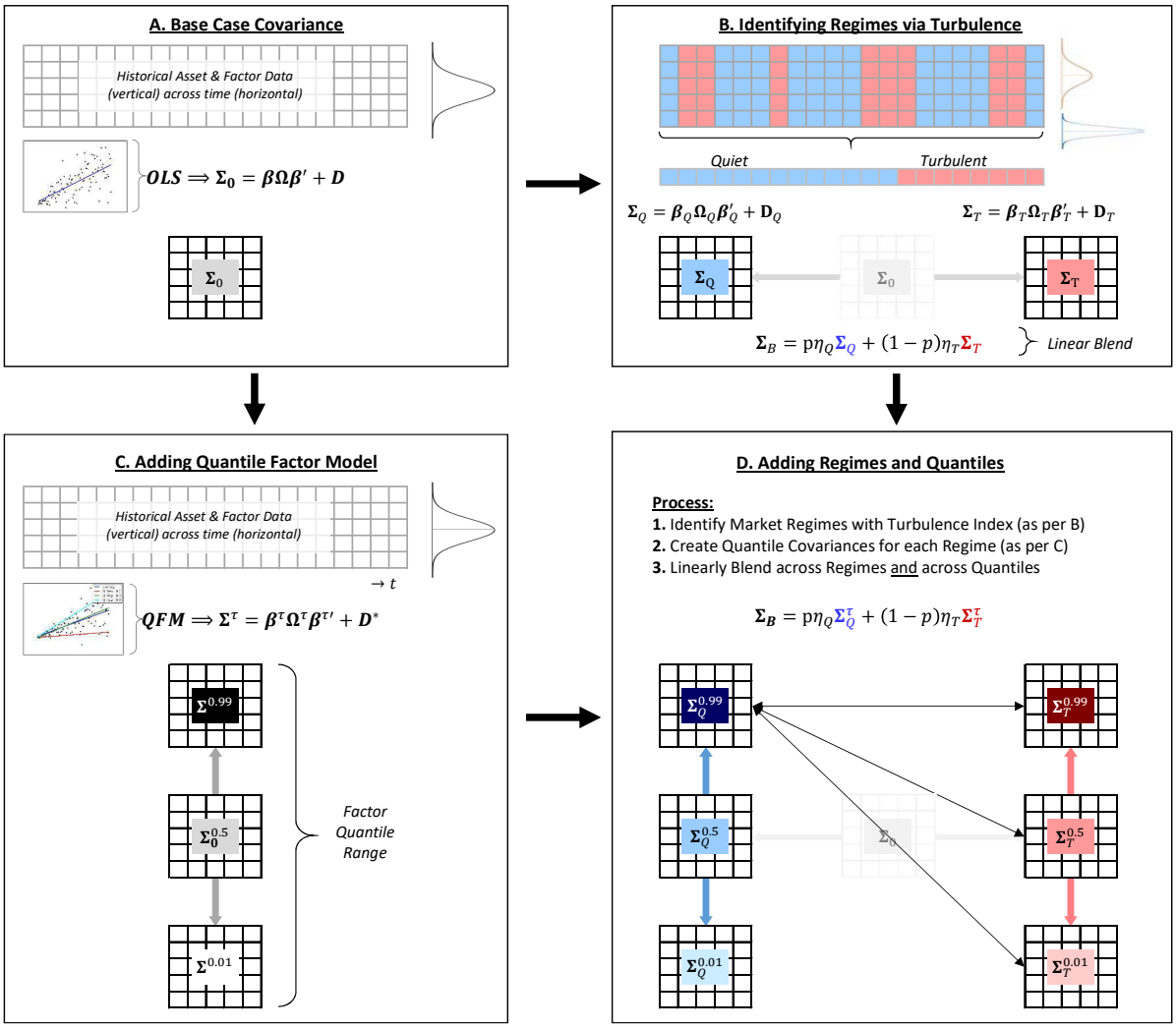


Figure 7.1. Stylised depiction showing how market regimes and quantile factor models may be incorporated in the covariance estimation procedure.

Finally, panel D shows how to incorporate both market regimes and quantile factor models into the covariance estimation procedure. Given our specific regime-switching model, and in the interest of simplicity, we suggest first identifying and partitioning the market data into regimes using the turbulence index process depicted in panel B. Quantile factor models should then be fitted separately to each regime in order to obtain two sets of quantile-specific covariance matrices. The final step then entails estimating the covariance matrix for input into the risk budgeting problem using the linear regime blending function, but where the two covariance matrices being blended are selected based on the specified quantiles of interest within each regime.

7.3 Extended Risk Budgeting: A South African Case Study

Having considered the theory and implementation of our framework, we now showcase its empirical application with a simple case study. The data for the case study consists of weekly returns for a set of South African stocks listed on the Johannesburg Stock Exchange (JSE) over the 10-year period between January 2008 to December 2017. The test set consists of the 40 largest stocks by index weight in the FTSE/JSE All Share Index (ALSI) as at 29 Dec 2017 for which a complete 10-year price history is available. From this set, we construct a market proxy portfolio based on the ALSI index weights of the selected stocks, normalised to sum to one. We make use of an illustrative risk factor model motivated by the work of Treynor and Mazuy (1966) and comprising two risk factors, namely, market portfolio returns and squared market portfolio returns.

For the case study, we restrict our attention to just two practical applications: scenario analysis and parameter uncertainty analysis. In both applications, the portfolio of interest is defined as the global minimum variance portfolio (GMV), constrained such that at least 30 stocks have a weight larger than 0.05%.[3] Scenario analysis and parameter uncertainty analysis both require a set of blended covariance matrices. This set, denoted $\mathcal{C}$, is created using Equation 7.13, where $p = \{0, 0.1, \ldots, 1\}$, $\eta = \{1, 1\}$, $\tau_Q = \{0.05, 0.10, \ldots, 0.95\}$ and $\tau_T = \{0.05, 0.10, \ldots, 0.95\}$.[4] Thus, our complete set $\mathcal{C}$ comprises 3971 matrices, covering the range of possible regime and quantile combinations in fairly fine granularity.

7.3.1 Scenario Analysis

The idea behind scenario analysis is to use covariance set $\mathcal{C}$ to determine how a given portfolio might behave under different regimes and in different areas of the multivariate asset distribution. Each matrix within the set $\mathcal{C}$ represents a possible future scenario that can be used to create a corresponding set of possible portfolio metrics. Take the example of a constrained GMV portfolio, w_{GMV}, calculated from the basic sample covariance matrix Σ_0. This portfolio has a volatility of $\sigma_0 = 14.3\%$. If, however, we consider that the sample covariance represents only one possible future scenario, and likely not the expected scenario, then it may be useful to

[3] This is achieved via the constraint functions in the optimisation problem 7.1, giving an approximate cardinality-constrained portfolio without the need for solving a much harder mixed integer optimisation problem.

[4] Equal risk-aversion parameters are selected for quiet and turbulent regimes to simplify the exposition of the case study results.

calculate the full range of possible portfolio volatilities $\mathcal{V}$ that this particular portfolio may take:

$$\mathcal{V}\left(w_{GMV}, \mathcal{C}\right) = \sqrt{w_{GMV}\mathcal{C}w'_{GMV}}. \tag{7.15}$$

The values in the set $\mathcal{V}$ can then be analysed relative to the original portfolio volatility σ_0, thereby enhancing our understanding of how regimes and quantiles affect the portfolio in terms of local optimality as well as portfolio robustness. Figure 7.2 plots the change in portfolio volatility when we move from a particular quantile in the quiet regime to the equivalent quantile in the turbulent regime. Given the nature of the two market regimes, we would expect the portfolio to have higher volatility during an increasingly turbulent regime, regardless of the quantile under consideration. This is confirmed by the graph, which shows that portfolio volatility increases by between 4% and 10% in all five quantiles considered when they move from a quiet regime to a turbulent one.

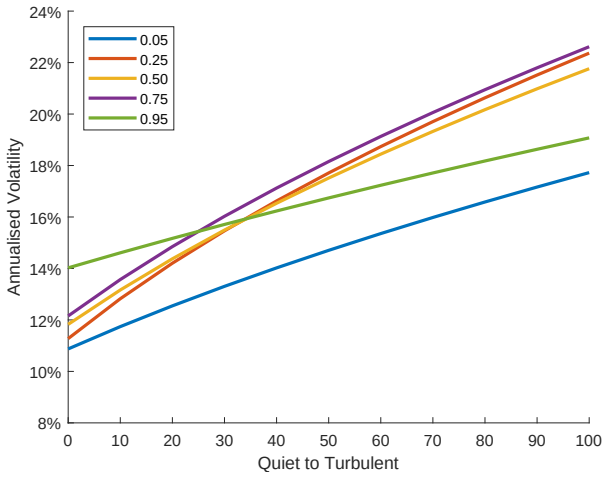


Figure 7.2. Profiles of w_{GMV} portfolio volatility across regimes for a selection of quantiles.

In contrast, the results for the individual quantiles are somewhat surprising. Firstly, as Cont (2001) and numerous others have shown, there is an asymmetric relationship between volatility and returns. Secondly, returns distributions are fat tailed. What these two facts suggest is that volatility should increase in periods when returns are negative and also when returns are large, which is the expected returns behavior in extreme quantiles. However, Figure 7.2 suggests the opposite. The covariance matrices estimated around the center of the conditional distribution, that is, quantiles $\tau = \{0.25, 0.50, 0.75\}$, generally produce higher volatility profiles than the extreme quantiles $\tau = \{0.05, 0.95\}$. In particular, the lowest quantile actually produces the lowest volatility across all regime blends, whereas one would perhaps expect the opposite.

This counter-intuitive result stems from the quantile covariance estimation procedure. Recall from Equation 7.12 that if the factor set remains constant, each quantile-

specific covariance matrix differs in the variation in factor betas alone. Although estimated using quantile regression and not OLS, these betas can still be interpreted as some measure of volatility-weighted asset-to-factor correlation. Ang and Chen (2002) showed that empirical equity correlations and volatilities are asymmetric across the return distribution, and that both the level and scale of this asymmetry are stock specific. Thus, it would appear that for the South African market, the estimated quantile factor betas are such that portfolio volatilities are actually lower in extremely negative return markets across all regimes as well as in extremely positive return markets for increasingly turbulent regimes. We leave further analysis into the underlying correlation and volatility components for future research.

We now turn our attention to the change in portfolio volatility across quantiles for a selection of six different regime blends, as displayed in Figure 7.3. These blends are obtained by starting with a regime that is 100% quiet and moving to one that is 100% turbulent in increments of 20%. The bounds on portfolio volatility are readily apparent from the figure. During a purely quiet regime, portfolio volatility is expected to range between 11% and 13%, whereas during a purely turbulent regime volatility varies between 18% and 23%. The impact of the conditional distribution quantiles is thus higher in more turbulent regimes.

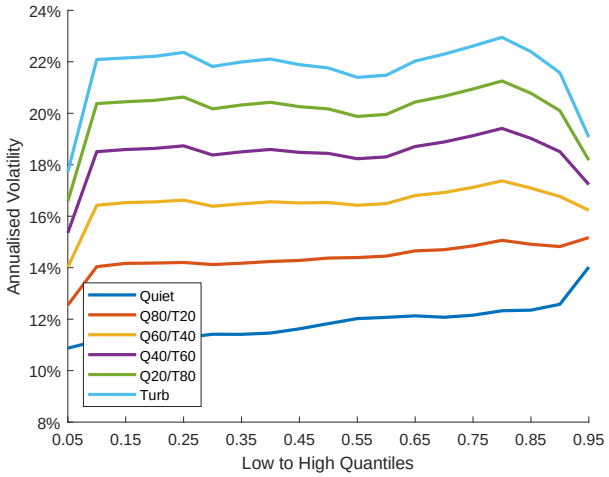


Figure 7.3. Profiles of w_{GMV} portfolio volatility across quantiles for a selection of regime blends.

Looking at the shape of the volatility profiles, the quiet regime shows an increase across the quantiles that is almost linear, while the turbulent regime profile is more erratic. In particular, the volatilities are relatively lower for the central quantiles and increase towards the edges, peaking at the 25^{th} and 85^{th} quantiles respectively. As one moves into the extreme quantiles though, the volatilities begin to drop, which is, again, somewhat counter-intuitive. As noted previously, this stems from the specific behavior of the quantile factor betas.

In summary, scenario analysis shows the potential shortcomings of using a single covariance matrix for portfolio construction. It also identifies bounds on the portfolio metrics during stressful market periods and with returns in the extremes of the distribution.

7.3.2 Parameter Uncertainty Analysis

The framework can also be used to test whether the estimated weights of a particular risk-optimal portfolio remain consistent across regimes and quantiles. For example, the GMV portfolio estimated from the sample covariance matrix should hopefully be similar to GMV portfolios estimated from covariance matrices in both quiet and turbulent regimes. Differences in these estimates can indicate parameter uncertainty.

To conduct the parameter uncertainty analysis, we use the extended risk budgeting optimisation given in Equation 7.14 and the covariance set $\mathcal{C}$ to generate a corresponding set of GMV portfolios, denoted by $\mathcal{W}$. Two methods can now be used to test for parameter uncertainty. Firstly, the weights of the portfolios in $\mathcal{W}$ can be compared to each other. Secondly, the regime- and quantile-specific minimum portfolio volatilities can be compared with the volatility profiles calculated from the base case portfolio, w_{GMV}.

Figure 7.4 shows the weight ranges for each stock across the optimal portfolio set $\mathcal{W}$. The interquartile range of the weights can be up to 3%, with the full range as large as 8% for some stocks, indicative of significant uncertainty in the optimal portfolio estimates.

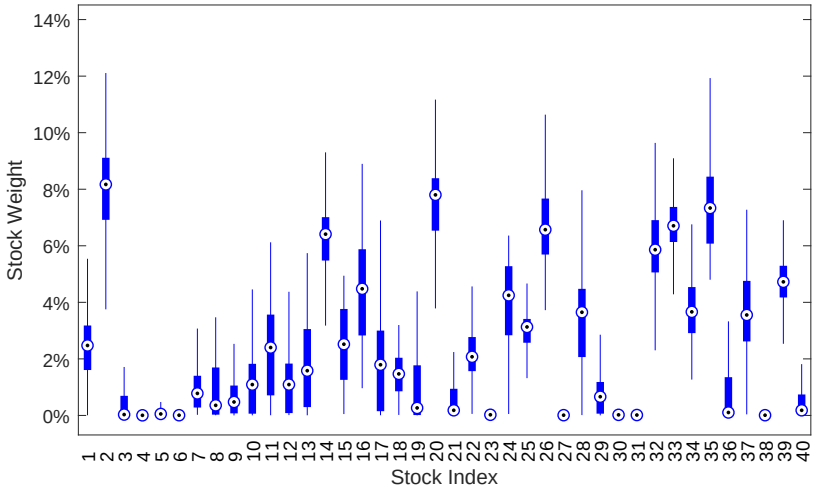


Figure 7.4. Asset weight dispersion of $\mathcal{W}$.

Figure 7.5 plots the difference in volatility between w_{GMV} and a selection of alternative GMV portfolios from $\mathcal{W}$ across a range of regime- and quantile-specific market scenarios. From the positive values in the graph, it is clear that the volatility of w_{GMV} is almost always higher than the corresponding volatilities of in the $\mathcal{W}$ portfolios for each of the market scenarios, and that the difference is particularly acute during turbulent regimes and extreme quantiles. Indeed, Figures 7.4 and 7.5 both suggest that the estimation of the optimal GMV portfolio is significantly influenced by regimes and quantiles.

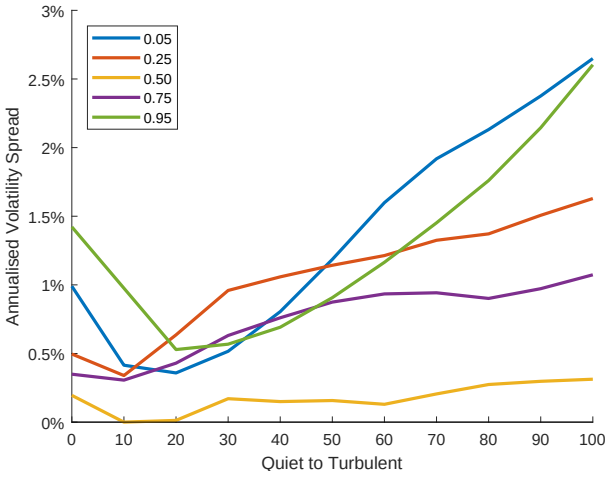


Figure 7.5. Differences in volatility between w_{GMV} and a selection of GMV portfolios from $\mathcal{W}$.

Figure 7.5 also confirms that we are able to recover the sample covariance matrix at a particular regime blend point on the median quantile profile. The volatility spread is approximately zero for the values $p = \{0.1, 0.2\}$, confirming that these regime- and quantile-specific GMV portfolios are approximately equal to w_{GMV}.

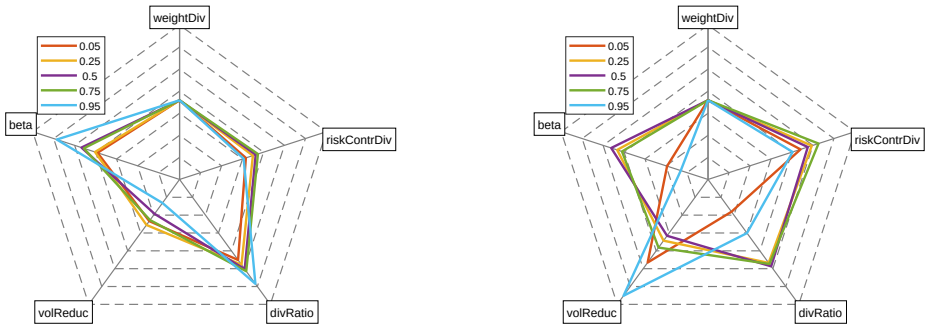
7.4 Multi-Objective and Multi-Scenario Optimisation via Risk Maps

In this section, we consider the issue of creating multi-objective and multi-scenario risk-optimal portfolios. When Richard and Roncalli (2015) introduced their generalised risk budgeting framework, they made use of spider plots (or radar plots) in order to graphically illustrate the behavior of a portfolio considered optimal in a single risk dimension across a number of other risk and diversification dimensions. In this section, we formalise this graphical approach to measuring multi-dimensional risk optimality through the concept of Risk Maps.

7.4.1 Introducing Risk Maps

A Risk Map is defined as a spider plot that displays a number of portfolio risk measures on a single graph. Each axis, or spoke, of the spider plot displays a different risk measure of the portfolio, with the outer-most point representing the optimal value achievable in that risk dimension and the centre point representing the least optimal point. Importantly, these optimal risk axis bounds need to be estimated, giving us the freedom to customise each Risk Map to the asset universe and constraint set of the given portfolio, or to suit a more general investment mandate. Once these bounds have been set, the portfolio will have a single point that plots on each of the risk axes, representing its relative optimality within each dimension. These points are then joined together to create a closed, irregular polygon; this represents the portfolio's Risk Map. The larger the surface area of the portfolio's Risk Map, the closer to optimal it is across the multiple risk dimensions specified.

Figure 7.6 displays the Risk Maps for quantile-specific GMV portfolios under quiet and turbulent market regimes. The axes chosen represent five common risk dimensions used in risk budgeting, each of which has a corresponding risk-optimal portfolio that can be calculated from the optimisation parameters given in Table 7.1.[5] Starting from the highest point and moving clockwise, the axes represent a portfolio's weight diversification (maximised by EW), risk contribution diversification (maximised by ERC), diversification ratio (maximised by MPD), volatility reduction (relative to the market portfolio, maximised by GMV), and portfolio beta (relative to the market portfolio beta of 1).



(a) Quiet regime Risk Map **(b)** Turbulent regime Risk Map

Figure 7.6. Risk Maps for quantile-specific GMV portfolios under quiet and turbulent market regimes.

[5]The corresponding minimum axis values are estimated through optimisation using a variant of Equation 7.14 and are specific to the asset universe and constraint set used throughout the case study.

Multi-objective optimality is defined as the total surface area of the polygon and is calculated by summing the areas between each of the axes. These areas are triangular in shape, formed by a common origin point and the risk values on each set of adjacent axes. We can readily interpret these risk points as vectors, and thus the area of each triangle A_Δ is equal to half the cross product of the two relevant risk vectors, r_i and r_j:

$$A_\Delta = \frac{1}{2}|r_i \times r_j| = \frac{1}{2}|r_1||r_2|sin(\theta), \tag{7.16}$$

where θ is the angle between the two vectors. Given that the plotted points on the Risk Map correspond to the magnitude of each vector and the relevant angle of each triangle is equal to 360° divided by the number of dimensions in the spider plot, the total area can easily be computed as

$$A = \frac{1}{2}\sin\left(\frac{360}{n}\right)\left(|r_1||r_2| + |r_2||r_3| + \cdots + |r_{n-1}||r_n| + |r_n||r_1|\right). \tag{7.17}$$

The area of the Risk Map is therefore a tractable measure of multi-objective risk optimality and can easily be tailored to accommodate any set of bounded portfolio measures. Investor preference as regards the risk dimensions can also be included, either by adding scaling factors to each of the risk vector magnitudes or by altering the orientation of the axes in the graph. The former option is considerably easier to implement, although the latter does have the advantage of being able to account directly for correlations between risk dimensions. We leave further analysis of preference-tailored Risk Maps to future research.

7.4.2 Regime- and Quantile-Optimal GMV Portfolios

In addition to multi-objective optimality, we can use Risk Maps and our extended risk budgeting framework to calculate portfolios that are multi-scenario optimal. In a similar vein to Section 7.3.1, we consider the goal of finding the global minimum variance portfolio; however, the minimum is now measured across a range of possible future scenarios. As a simple example, we calculate two portfolios, denoted w_T and w_Q, and consider whether these portfolios are better estimates of a multi-scenario GMV portfolio than the original w_{GMV} portfolio estimated from the sample covariance matrix. The two candidate portfolios w_T and w_Q are obtained by averaging the optimal asset weights in two particular subsets of $\mathcal{W}$, denoted as $\mathcal{W}_T$ and $\mathcal{W}_Q$ respectively. The turbulent subset $\mathcal{W}_T$ includes a selection of scenarios for which the regime probability $p = \{0.8, 0.9, 1\}$, while the quiet subset $\mathcal{W}_Q$ includes scenarios for which $p = \{0, 0.1, 0.2\}$. For each regime probability value, we include quantile scenarios in the set $\tau = \{0.05, 0.1, \ldots, 0.95\}$. Only equivalent regime quantile blends are included in the two sets though, meaning that $\tau_T = \tau_Q$ always. The

final multi-scenario portfolios w_T and w_Q are then calculated as the average asset weights across each of the 57 GMV-optimal portfolio subsets respectively.

Table 7.2 gives the volatility reduction of the three portfolios relative to the market portfolio across 11 market scenarios; namely, the full sample scenario combined with two sets of five quantile scenarios from the quiet and turbulent regimes. Looking first at the sample covarance scenario, we see that portfolio w_{GMV} achieves a volatility reduction of 24.7% compared to the market portfolio. It also performs slightly better than the competing portfolios w_T and w_Q, which provide volatility reductions of 22.6% and 23.0% respectively. This is to be expected, given that portfolio w_{GMV} is by definition risk-optimal under the sample covariance scenario.

Table 7.2. Volatility reduction of three competing GMV portfolios in eleven market scenarios.

Market Scenario		**Volatility Reduction (%)**		
Regime	Quantile	w_{GMV}	w_T	w_Q
Sample	*n.a.*	24.7	22.6	23.0
	5^{th}	25.6	21.7	27.8
	25^{th}	26.0	22.8	26.9
Quiet	50^{th}	22.5	19.6	22.9
	75^{th}	24.2	21.4	25.1
	95^{th}	20.1	16.6	22.1
	5^{th}	30.3	33.3	25.8
	25^{th}	24.9	27.1	22.7
Turbulent	50^{th}	27.1	25.8	23.1
	75^{th}	28.1	29.1	26.2
	95^{th}	41.7	42.6	39.3

Comparing the volatility reduction across the regime-specific scenarios, we see that both the competing portfolios perform better than w_{GMV} in their respective regimes; again, as expected. The greatest improvement is seen for the w_T portfolio in the turbulent regime. The differences, however, are not large enough for any one of the three portfolios to be considered a clear winner across all scenarios. Even so, if one makes the reasonable assumption that we would rather sacrifice some volatility reduction during quiet regimes in order to have maximum volatility reduction during turbulent regimes, then the portfolio w_T becomes quite appealing.

One potential caveat bears mentioning: at what cost do we obtain the additional volatility reduction in portfolios w_T and w_Q with respect to the remaining risk dimensions? To answer this, Table 7.3 gives the area of the Risk Maps for each of the three portfolios across the 11 scenarios described previously. Looking at the the sample covariance scenario, we see that the w_{GMV} portfolio covers an area of 28%, while the competing multi-scenario portfolios cover areas of 31% and 33% respectively. This means that although portfolio w_{GMV} provides a slightly better

reduction in volatility compared with the other candidate portfolios, it is relatively sub-optimal across the remaining risk dimensions.

Table 7.3. Risk Map areas of three competing GMV portfolios in eleven market scenarios.

Market Scenario		**Risk Map Area (%)**		
Regime	Quantile	w_{GMV}	w_T	w_Q
Sample	n.a.	28	31	33
	5^{th}	22	24	28
	25^{th}	25	27	31
Quiet	50^{th}	25	27	31
	75^{th}	27	29	32
	95^{th}	25	27	33
	5^{th}	18	20	23
	25^{th}	30	36	34
Turbulent	50^{th}	31	35	35
	75^{th}	31	38	37
	95^{th}	20	24	25

Considering the remainder of Table 7.3, we find that portfolios w_T and w_Q both display consistently superior Risk Map areas across all scenarios compared with w_{GMV}. This means that the superior multi-scenario volatility reduction does not come at a significant cost in terms of the total risk. Moreover, taken in conjunction with the results given in Table 7.2, this observation strengthens the case for selecting portfolio w_T as the optimal multi-scenario GMV portfolio, since it offers improved volatility reduction during turbulent regimes as well as a superior total risk profile across all regimes.

7.5 Conclusion

In this chapter, we combine several disparate avenues in this area of literature to create a novel, unified risk-based optimisation framework. Specifically, we extend the existing risk budgeting approach of Richard and Roncalli (2015) to allow for changing market regimes, factor dependence and a nonlinear and asymmetric market structure. We show that the existing framework can be readily extended to include a factor-dependent return process using the standard models available in the literature. Structural changes in market conditions are incorporated into the framework using a regime-switching turbulence index, while quantile factor models allow us to account for the nonlinear and asymmetric market dependence structure. Most importantly, the extended framework is only comprised of a series of linear models and is thus simple to understand and implement.

We discuss the theory of our extended framework at length and also provide details of how one might implement it in practice. The framework is general enough to

enable the practitioner to focus on four cases of covariance estimation: the base case from the full data sample, a regime-switching estimation, a quantile-specific estimation, and an estimation that incorporates both regimes and quantiles.

By means of a South African case study, we demonstrate two empirical applications of the framework: scenario analysis and parameter uncertainty analysis. Via scenario analysis, we demonstrate that the volatility of the GMV portfolio estimated from the full data sample differs meaningfully across both regimes and quantiles. This analysis sets our expectation of how a given GMV portfolio is likely to behave under different market conditions. In our parameter uncertainty analysis, we generate estimates of the GMV portfolio that are specific to the selected market regime and return distribution quantile. The effect of model parameter uncertainty can then be quantified by the variation in the estimated weights and the volatility reduction performance of the portfolios. We find empirical evidence of significant variation in both measures.

Finally, we introduce the concept of Risk Maps, which provide us with a graphical approach for estimating and evaluating multi-objective and multi-scenario risk-optimal portfolios. We use Risk Maps empirically to quantify the optimality of three candidate GMV portfolios across multiple risk dimensions and market scenarios. We show that the GMV portfolio estimated from multiple turbulent regime scenarios provides improved risk reduction across almost all market scenarios relative to the full sample GMV portfolio, and, furthermore, that this portfolio displays superior optimality across the remaining risk dimensions.

There exist several avenues for further research based on the work in this chapter. Firstly, one could examine how the betas of South African stocks change across regime blends and return quantiles. Some of the results given above suggest that these betas may exhibit interesting nonlinear behaviour at the extreme quantiles. Secondly, it would be practically useful to backtest the performance of standard risk-based optimal portfolios and compare them to their multi-scenario optimal counterparts. Lastly, there is scope for considerable further research to be done – theoretical and empirical – on the application of Risk Maps across the portfolio and risk management spectrum.

Bibliography

Aıt-Sahalia, Yacine and Jefferson Duarte (2003). „Nonparametric Option Pricing under Shape Restrictions“. In: *Journal of Econometrics* 116.1, pp. 9–47.

Aït-Sahalia, Yacine and Andrew W Lo (1998). „Nonparametric Estimation of State-price Densities implicit in Financial Asset Prices“. In: *The Journal of Finance* 53.2, pp. 499–547.

Alvarez-Ramirez, Jose, Myriam Cisneros, Carlos Ibarra-Valdez, and Angel Soriano (2002). „Multifractal Hurst analysis of crude oil prices“. In: *Physica A: Statistical Mechanics and its Applications* 313.3-4, pp. 651–670.

Alvarez-Ramirez, Jose, Jesus Alvarez, Eduardo Rodriguez, and Guillermo Fernandez-Anaya (2008). „Time-varying Hurst exponent for US stock markets“. In: *Physica A: Statistical Mechanics and its Applications* 387.24, pp. 6159–6169.

Amenc, Noel, Romain Deguest, Felix Goltz, et al. (2014). *Risk allocation, factor investing and smart beta: Reconciling innovations in equity portfolio construction*. Tech. rep. EDHEC-Risk Institute.

Andersen, Torben G and Tim Bollerslev (1997). „Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns“. In: *The journal of Finance* 52.3, pp. 975–1005.

Ang, Andrew (2014). *Asset management: A systematic approach to factor investing*. Oxford University Press.

Ang, Andrew and Geert Bekaert (2002). „International asset allocation with regime shifts“. In: *The Review of Financial Studies* 15.4, pp. 1137–1187.

Ang, Andrew and Joseph Chen (2002). „Asymmetric correlations of equity portfolios“. In: *Journal of financial Economics* 63.3, pp. 443–494.

Ang, Andrew and Allan Timmermann (2012). „Regime changes and financial markets“. In: *Annu. Rev. Financ. Econ.* 4.1, pp. 313–337.

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang (2006). „The Cross-Section of Volatility and Expected Returns“. In: *The Journal of Finance* 61.1, pp. 259–299.

Antonacci, Gary (2013). „Absolute Momentum: A Simple Rule-Based Strategy and Universal Trend-Following Overlay“. In: *SSRN Electronic Journal* 2244633.

Arnott, Robert, Noah Beck, and Vitali Kalesnik (2016). „Timing 'Smart Beta' Strategies? Of Course! Buy Low, Sell High!“ In: *SSRN Electronic Journal* 3040956.

Asness, Cliff and Andrea Frazzini (2013). „The devil in HML's details". In: *Journal of Portfolio Management* 39.4, pp. 49–68.

Asness, Clifford, Andrea Frazzini, Ronen Israel, and Tobias Moskowitz (2015). „Fact, Fiction, and Value Investing". In: *The Journal of Portfolio Management* 42, pp. 34–52.

Asness, Clifford, Swati Chandra, Antti Ilmanen, and Ronen Israel (2017). „Contrarian Factor Timing is Deceptively Difficult". In: *The Journal of Portfolio Management* 43.5, pp. 72–87. eprint: `https://jpm.iijournals.com/content/43/5/72.full.pdf`.

Asness, Clifford S., Tobias J. Moskowitz, and Lasse Heje Pedersen (2013). „Value and Momentum Everywhere". In: *The Journal of Finance* 68.3, pp. 929–985.

Asness, Clifford S, Andrea Frazzini, and Lasse Heje Pedersen (2014). „Quality minus junk". In: *Review of Accounting Studies*, pp. 1–79.

Atkins, Allen B and Pin T Ng (2014). „Refining our understanding of beta through quantile regressions". In: *Journal of Risk and Financial Management* 7.2, pp. 67–79.

Audrino, Francesco, Robert Huitema, and Markus Ludwig (2015). „An Empirical Analysis of the Ross Recovery Theorem". In: *SSRN Electronic Journal* 2433170.

Avellaneda, Marco (2009). „Dispersion Trading". In: *Courant Institute of Mathematical Sciences: Mathematics in Finance lectures*.

Backwell, Alex (2015). „State Prices and Implementation of the Recovery Theorem". In: *Journal of Risk and Financial Management* 8.1, pp. 2–16.

Baillie, Richard T, Tim Bollerslev, and Hans Ole Mikkelsen (1996). „Fractionally integrated generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 74.1, pp. 3–30.

Baker, Malcolm, Brendan Bradley, and Ryan Taliaferro (2014). „The low-risk anomaly: A decomposition into micro and macro effects". In: *Financial Analysts Journal* 70.2, pp. 43–58.

Basiewicz, PG and CJ Auret (2009). „Another look at the cross-section of average returns on the JSE". In: *Investment Analysts Journal* 38.69, pp. 23–38.

– (2010). „Feasibility of the Fama and French three factor model in explaining returns on the JSE". In: *Investment Analysts Journal* 39.71, pp. 13–25.

Bassett Jr, Gilbert W and Hsiu-Lang Chen (2001). „Portfolio style: Return-based attribution using quantile regression". In: *Empirical Economics* 26.1, pp. 293–305.

Bates, David S (2000). „Post-'87 Crash Fears in the S&P 500 Futures Option Market". In: *Journal of Econometrics* 94.1, pp. 181–238.

Baule, Rainer, Olaf Korn, and Sven Saßning (2015). „Which Beta Is Best? On the Information Content of Option-Implied Betas". In: *European Financial Management*.

Beasley, John E. (2013). „Portfolio optimisation: models and solution approaches". In: *Theory Driven by Influential Applications*. Informs, pp. 201–221.

Bhansali, Vineer, Josh Davis, Graham Rennison, Jason Hsu, and Feifei Li (2012). „The Risk in Risk Parity: A Factor-Based Analysis of Asset-Based Risk Parity". In: *The Journal of Investing* 21.3, pp. 102–110.

Biagini, Francesca, Yaozhong Hu, Bernt Øksendal, and Tusheng Zhang (2008). *Stochastic calculus for fractional Brownian motion and applications*. Springer Science & Business Media.

Black, Fischer (1972). „Capital market equilibrium with restricted borrowing". In: *The journal of business* 45.3, pp. 444–455.

– (1976). „The Pricing of Commodity Contracts". In: *Journal of Financial Economics* 3.1, pp. 167–179.

Black, Fischer and Myron Scholes (1973). „The Pricing of Options and Corporate Liabilities". In: *The Journal of Political Economy*, pp. 637–654.

Bliss, Robert R and Nikolaos Panigirtzoglou (2002). „Testing the Stability of Implied Probability Density Functions". In: *Journal of Banking & Finance* 26.2, pp. 381–422.

– (2004). „Option-Implied Risk Aversion Estimates". In: *The Journal of Finance* 59.1, pp. 407–446.

Blitz, David and Pim Van Vliet (2007). „The volatility Effect". In: *Journal of Portfolio Management* 34.1, pp. 102–113.

Borovička, Jaroslav, Lars Peter Hansen, and José A Scheinkman (2016). „Misspecified Recovery". In: *The Journal of Finance*.

Borovkova, SA, FJ Permana, and JAM Van Der Weide (2012). „American basket and spread option pricing by a simple binomial tree". In: *Journal of Derivatives* 19.4, p. 29.

Breeden, Douglas T and Robert H Litzenberger (1978). „Prices of State-Contingent Claims implicit in Option Prices". In: *Journal of Business*, pp. 621–651.

Brinson, Gary P, L Randolph Hood, and Gilbert L Beebower (1995). „Determinants of portfolio performance". In: *Financial Analysts Journal* 51.1, pp. 133–138.

Britten-Jones, Mark and Anthony Neuberger (2000). „Option prices, implied price processes, and stochastic volatility". In: *The Journal of Finance* 55.2, pp. 839–866.

Buchen, Peter W and Michael Kelly (1996). „The Maximum Entropy Distribution of an Asset inferred from Option Prices". In: *Journal of Financial and Quantitative Analysis* 31.01, pp. 143–159.

Burns, Arthur F. and Wesley C. Mitchell (1946). *Measuring Business Cycles*. National Bureau of Economic Research, Inc.

Buss, Adrian and Grigory Vilkov (2012). „Measuring Equity Risk with Option-Implied Correlations". In: *Review of Financial Studies* 25.10, pp. 3113–3140.

Cajueiro, Daniel O and Benjamin M Tabak (2004). „The Hurst exponent over time: testing the assertion that emerging markets are becoming more efficient". In: *Physica A: Statistical Mechanics and its Applications* 336.3-4, pp. 521–537.

Campbell, John Y. and Robert J. Shiller (1988). „Stock Prices, Earnings, and Expected Dividends". In: *The Journal of Finance* 43.3, pp. 661–676.

Canakgoz, Nilgun A and John E Beasley (2009). „Mixed-integer programming approaches for index tracking and enhanced indexation". In: *European Journal of Operational Research* 196.1, pp. 384–399.

Carhart, Mark M (1997). „On persistence in mutual fund performance“. In: *The Journal of finance* 52.1, pp. 57–82.

Carr, Peter and Dilip B Madan (2005). „A Note on Sufficient Conditions for No Arbitrage“. In: *Finance Research Letters* 2.3, pp. 125–130.

Carr, Peter and Jiming Yu (2012). „Risk, Return, and Ross Recovery“. In: *Journal of Derivatives* 20.1, p. 38.

Carr, Peter, Hélyette Geman, Dilip B Madan, and Marc Yor (2003). „Stochastic Volatility for Lévy Processes“. In: *Mathematical Finance* 13.3, pp. 345–382.

– (2005). „Pricing options on realized variance“. In: *Finance and Stochastics* 9.4, pp. 453–475.

Cazalet, Zélia and Thierry Roncalli (2014). „Facts and Fantasies About Factor Investing“. In: *SSRN Electronic Journal* 2524547.

Cesarone, Francesco, Jacopo Moretti, and Fabio Tardella (2014). „Does greater diversification really improve performance in portfolio selection?“ In: *SSRN Electronic Journal* 2473630.

Chen, Liang, Juan J. Dolado, and Jesús Gonzalo (2016). „Quantile factor models“. In: *SSRN Electronic Journal* 3126210. Working Paper.

Chow, George, Eric Jacquier, Mark Kritzman, and Kenneth Lowry (1999). „Optimal portfolios in good times and bad“. In: *Financial Analysts Journal* 55.3, pp. 65–73.

Clarke, Roger, Harindra De Silva, and Steven Thorley (2002). „Portfolio constraints and the fundamental law of active management“. In: *Financial Analysts Journal* 58.5, pp. 48–66.

Cochrane, John H (2001). *Asset Pricing*. Princeton University Press.

– (2011). „Presidential address: Discount rates“. In: *The Journal of finance* 66.4, pp. 1047–1108.

Cohen, Tim, Brian Leite, Darby Nelson, and Andy Browder (2014). „Active share: A misunderstood measure in manager selection“. In: *Fidelity Leadership Series Investment Insights (February)*.

Comte, Fabienne and Eric Renault (1998). „Long memory in continuous-time stochastic volatility models“. In: *Mathematical finance* 8.4, pp. 291–323.

Comte, Fabio and Eric Renault (1996). „Long memory continuous time models“. In: *Journal of Econometrics* 73.1, pp. 101–149.

Cont, Rama (2001). „Empirical properties of asset returns: stylized facts and statistical issues“. In: *Quantitative Finance* 1 (2), pp. 223–236.

Cont, Rama, José Da Fonseca, et al. (2002). „Dynamics of implied volatility surfaces“. In: *Quantitative finance* 2.1, pp. 45–60.

Cooper, Neil (1999). „Testing Techniques for Estimating Implied RNDs from the Prices of European-Style Options“. In: *Proceedings of the BIS workshop on implied PDFs, Bank of International Settlements, Basel*.

Coutant, Sophie, Eric Jondeau, and Michael Rockinger (1998). *Reading Interest Rate and Bond Futures Options' Smiles: How PIBOR and Notional Operators Appreciated the 1997 French Snap Election*. Tech. rep. Banque de France.

Cox, John C and Stephen A Ross (1976). „The Valuation of Options for Alternative Stochastic Processes“. In: *Journal of Financial Economics* 3.1-2, pp. 145–166.

Cremers, K. J. Martijn and Antti Petajisto (2009). „How Active Is Your Fund Manager? A New Measure That Predicts Performance“. In: *The Review of Financial Studies* 22.9, pp. 3329–3365. eprint: `http://oup.prod.sis.lan/rfs/article-pdf/22/9/3329/24430249/hhp057.pdf`.

Damghani, Babak Mahdavi and Andrew Kos (2013). „De-arbitraging With a Weak Smile: Application to Skew Risk“. In: *Wilmott* 2013.64, pp. 40–49.

De Franco, Carmine, Bruno Monnier, Johann Nicolle, and Ksenya Rulik (2016). „How Different Are Alternative Beta Strategies?“ In: *The Journal of Index Investing* 7, pp. 57–77.

De Marco, Stefano and Claude Martini (2009). „Quasi-Explicit Calibration of Gatheral SVI model“. In: *Zeliade Systems*.

De Silva, Harindra, Steven Thorley, and Roger Clarke (2006). „The Fundamental Law of Active Portfolio Management“. In: *Journal of Investment Management* 4.

Deguest, Romain, Lionel Martellini, and Attilio Meucci (2013). „Risk Parity and Beyond - From Asset Allocation to Risk Allocation Decisions“. In: *SSRN Electronic Journal* 2355778.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal (2007). „Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?“ In: *The review of Financial studies* 22.5, pp. 1915–1953.

DeMiguel, Victor, Yuliya Plyakha, Raman Uppal, and Grigory Vilkov (2013). „Improving Portfolio Selection Using Option-Implied Volatility and Skewness“. In: *Journal of Quantitative & Financial Analysis* 48.6, pp. 1813–1845.

Deng, Qian (2008). „Volatility Dispersion Trading“. In: *SSRN Electronic Journal* 1156620.

Derman, Emanuel (1996). „Outperformance options“. In: *I. Nelken, The handbook of exotic options: instruments, analysis, and application, McGraw-Hill*.

Derman, Emanuel, Iraj Kani, and Joseph Z Zou (1996). „The local volatility surface: Unlocking the information in index option prices“. In: *Financial analysts journal* 52.4, pp. 25–36.

Dubynskiy, Sergey and Robert S Goldstein (2013). „Recovering Drifts and Preference Parameters from Financial Derivatives“. In: *SSRN Electronic Journal* 2244394.

Dumas, Bernard, Jeff Fleming, and Robert E Whaley (1998). „Implied Volatility Functions: Empirical Tests“. In: *The Journal of Finance* 53.6, pp. 2059–2106.

Dupire, Bruno (1994). „Pricing with a smile“. In: *Risk* 7.1, pp. 18–20.

– (2006). „Model free results on volatility derivatives“. In: *SAMSI, Research Triangle Park*.

Elliott, Robert J and John Van Der Hoek (2003). „A general fractional white noise theory and applications to finance“. In: *Mathematical Finance* 13.2, pp. 301–330.

Embrechts, Paul, Filip Lindskog, and Alexander McNeil (2001). „Modelling dependence with copulas“. In: *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*.

Engle, Robert (2002). „Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models“. In: *Journal of Business & Economic Statistics* 20.3, pp. 339–350.

Engle, Robert F and Simone Manganelli (2004). „CAViaR: Conditional autoregressive value at risk by regression quantiles“. In: *Journal of Business & Economic Statistics* 22.4, pp. 367–381.

Faber, Mebane T. (2007). „A Quantitative Approach to Tactical Asset Allocation“. In: *The Journal of Wealth Management* 9.4, pp. 69–79. eprint: `https://jwm.iijournals.com/content/9/4/69.full.pdf`.

Fama, Eugene F. and Kenneth R. French (1993). „Common risk factors in the returns on stocks and bonds“. In: *Journal of Financial Economics* 33.1, pp. 3 –56.

– (2008). „Dissecting Anomalies“. In: *Journal of Finance* 63.4, pp. 1653–1678.

– (2015). „A five-factor asset pricing model“. In: *Journal of Financial Economics* 116.1, pp. 1 –22.

Figlewski, Stephen (2008). „Estimating the Implied Risk Neutral Density“. In: *Volatility and Time Series Econometrics*.

Fitzgibbons, Shaun, Jacques Friedman, Lukasz Pomorski, and Laura Serban (2017). „Long-Only Style Investing: Don't Just Mix, Integrate“. In: *The Journal of Investing* 26, pp. 153–164.

Flint, Emlyn and Eben Maré (2017a). „Estimating option-implied distributions in illiquid markets and implementing the Ross recovery theorem“. In: *South African Actuarial Journal* 17, pp. 1–28.

– (2017b). „Fractional Black-Scholes option pricing, volatility calibration and implied Hurst exponents in South African context“. In: *South African Journal of Economic and Management Sciences* 20.1, p. 11.

Flint, Emlyn and Simon du Plooy (2018). „Extending Risk Budgeting for Market Regimes and Quantile Factor Models“. In: *Journal of Investment Strategies*.

Flint, Emlyn, Florence Chikurunhe, and Anthony J. Seymour (2012). „(Un)Modelling the Volatility Surface: Valuing South African Volatility Surfaces Via Risk-Neutral Historic Return Distributions“. In: *SSRN Electronic Journal* 2767409.

Flint, Emlyn, Anthony Seymour, and Florence Chikurunhe (2013a). „Trends in the South African Hedging Space Part I“. In: *SSRN Electronic Journal* 2767416.

– (2013b). „Trends in the South African Hedging Space Part II“. In: *SSRN Electronic Journal* 2767421.

Flint, Emlyn, Florence Chikurunhe, and Anthony Seymour (2013c). *Understanding Rand-Hedge Stocks*. Tech. rep. Legae Peresec Research.

Flint, Emlyn, Anthony Seymour, and Florence Chikurunhe (2014). „Adapting to Market Regimes with Timed Hedging“. In: *SSRN Electronic Journal* 2767427.

– (2015a). „In Search of the Perfect Hedge“. In: *SSRN Electronic Journal* 2767532.

Flint, Emlyn, Florence Chikurunhe, and Anthony Seymour (2015b). „The Cost of a Free Lunch: Dabbling in Diversification“. In: *SSRN Electronic Journal* 2767436.

Flint, Emlyn, Anthony Seymour, and Florence Chikurunhe (2017a). „Factor Investing in South Africa“. In: *Alternative Investment Analyst Review* 6.2, pp. 19–36.

Flint, Emlyn, Florence Chikurunhe, and Anthony Seymour (2017b). „Regime-Based Tactical Allocation for Factor and Balanced Portfolios“. In: *SSRN Electronic Journal* 997410.

Frazzini, Andrea and Lasse Heje Pedersen (2014). „Betting against beta“. In: *Journal of Financial Economics* 111.1, pp. 1 –25.

Frazzini, Andrea, Jacques Friedman, and Lukasz Pomorski (2016). „Deactivating Active Share“. In: *Financial Analysts Journal* 72, pp. 1–8.

Gatheral, Jim (2004). „A Parsimonious Arbitrage-Free Implied Volatility Parameterization with Application to the Valuation of Volatility Derivatives“. In: *Presentation at Global Derivatives & Risk Management, Madrid*.

– (2011). *The Volatility Surface: A Practitioner's Guide*. Vol. 357. John Wiley & Sons.

Gatheral, Jim and Antoine Jacquier (2014). „Arbitrage-free SVI Volatility Surfaces“. In: *Quantitative Finance* 14.1, pp. 59–71.

Gray, Wesley R and Tobias E Carlisle (2012). *Quantitative Value: A Practitioner's Guide to Automating Intelligent Investment and Eliminating Behavioral Errors*. Vol. 836. John Wiley & Sons.

Gray, Wesley R and Jack R Vogel (2016). *Quantitative Momentum: A Practitioner's Guide to Building a Momentum-Based Stock Selection System*. John Wiley & Sons.

Grinold, Richard C. (1989). „The fundamental law of active management“. In: *The Journal of Portfolio Management* 15.3, pp. 30–37. eprint: `https://jpm.iijournals.com/content/15/3/30.full.pdf`.

Hagan, Patrick S, Deep Kumar, Andrew S Lesniewski, and Diana E Woodward (2002). „Managing Smile Risk“. In: *The Best of Wilmott*, p. 249.

Hamilton, James D. (1989). „A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle“. In: *Econometrica* 57.2, pp. 357–384.

Harvey, Campbell R., Yan Liu, and Heqing Zhu (2015). „... and the Cross-Section of Expected Returns“. In: *The Review of Financial Studies* 29.1, pp. 5–68.

Herfindahl, O. C. (1950). „Concentration in the US Steel Industry“. Ph. D. thesis. New York: Columbia University.

Heston, Steven L (1993). „A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency options“. In: *Review of Financial Studies* 6.2, pp. 327–343.

Hoffstein, Corey (2016). *Multi-Factor: Mix or Integrate?* Tech. rep. Newfound Research.

Homescu, Cristian (2014). „Many risks, one (optimal) portfolio“. In: *SSRN Electronic Journal* 2473776.

– (2015). „Better investing through factors, regimes and sensitivity analysis“. In: *SSRN Electronic Journal* 2557236.

Hu, Yaozhong and Bernt Øksendal (2003). „Fractional white noise calculus and applications to finance“. In: *Infinite dimensional analysis, quantum probability and related topics* 6.01, pp. 1–32.

Hunstad, Michael and Jordan Dekhayser (2015). „Evaluating the Efficiency of "Smart Beta" Indexes“. In: *The Journal of Index Investing* 6.1, pp. 111–121.

Hurst, Harold Edwin (1951). „Long-term storage capacity of reservoirs“. In: *Trans. Amer. Soc. Civil Eng.* 116, pp. 770–799.

Hwang, Soosung and Stephen Satchell (2001). „Tracking Error: Ex-Ante versus Ex-Post Measures“. In: *Journal of Asset Management* 2.

Jackwerth, Jens Carsten and Marco Menner (2018). „Does the Ross Recovery Theorem Work Empirically?“ In: *SSRN Electronic Journal* 2960733.

Jefferis, Keith and Pako Thupayagale (2008). „Long memory in Southern African stock markets“. In: *South African Journal of Economics* 76.3, pp. 384–398.

Jegadeesh, Narasimhan and Sheridan Titman (1993). „Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency“. In: *The Journal of Finance* 48.1, pp. 65–91. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1993.tb04702.x`.

Jensen, Christian Skov, David Lando, and Lasse Heje Pedersen (2018). „Generalized Recovery“. In: *SSRN Electronic Journal* 2674541.

Jondeau, Eric and Michael Rockinger (2000). „Reading the Smile: The Message Conveyed by Methods which Infer Risk Neutral Densities“. In: *Journal of International Money and Finance* 19.6, pp. 885–915.

Karuppiah, Jeyanthi and Cornelis A Los (2005). „Wavelet multiresolution analysis of high-frequency Asian FX rates, Summer 1997“. In: *International Review of Financial Analysis* 14.2, pp. 211–246.

Kempf, Alexander, Olaf Korn, and Sven Saßning (2015). „Portfolio Optimization using Forward-looking Information“. In: *Review of Finance* 19.1, pp. 467–490.

Khomo, Melvin and Meshach Aziakpono (2007). „Forecasting recession in South Africa: A comparison of the yield curve and other economic indicators“. In: *South African Journal of Economics* 75, pp. 194–212.

Kiriu, Takuya and Norio Hibiki (2015). „Estimating Forward Looking Distribution with the Ross Recovery Theorem“. In: *Keio University Working Paper*.

Kirk, E (1995). „Correlation in the energy markets.Managing energy price risk“. In: *Risk Publications*, pp. 71–78.

Koenker, Roger and Gilbert Bassett Jr (1978). „Regression quantiles“. In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Kostakis, Alexandros, Nikolaos Panigirtzoglou, and George Skiadopoulos (2011). „Market Timing with Option-implied Distributions: A Forward-looking Approach“. In: *Management Science* 57.7, pp. 1231–1249.

Kotzé, Antonie and Angelo Joseph (2009). „Constructing a South African Index Volatility Surface from Exchange Traded Data“. In: *SSRN Electronic Journal* 2198357.

Kotzé, Antonie, Coenraad CA Labuschagne, Merell L Nair, and Nadine Padayachi (2013). „Arbitrage-free implied volatility surfaces for options on single stock futures“. In: *The North American Journal of Economics and Finance* 26, pp. 380–399.

Kritzman, Mark and Yuanzhen Li (2010). „Skulls, Financial Turbulence, and Risk Management“. In: *Financial Analysts Journal* 66.5, pp. 30–41.

Kritzman, Mark, Sebastien Page, and David Turkington (2012). „Regime Shifts: Implications for Dynamic Strategies“. In: *Financial Analysts Journal* 68.3, pp. 22–39.

Laloux, Laurent, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud (2000). „Random matrix theory and financial correlations“. In: *International Journal of Theoretical and Applied Finance* 3.03, pp. 391–397.

Ledoit, Olivier and Michael Wolf (2004). „Honey, I shrunk the sample covariance matrix“. In: *The Journal of Portfolio Management* 30.4, pp. 110–119.

Lee, Roger W (2004). „The moment formula for implied volatility at extreme strikes“. In: *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 14.3, pp. 469–480.

Li, Kinrey and Rong Chen (2014). „Implied Hurst exponent and fractional implied volatility: A variance term structure model“. In: *SSRN Electronic Journal* 2383618.

Lo, Andrew (2004). „The Adaptive Markets Hypothesis“. In: *The Journal of Portfolio Management* 30.5, pp. 15–29. eprint: `https://jpm.iijournals.com/content/30/5/15.full.pdf`.

Lobosco, Angelo and Dan DiBartolomeo (1997). „Approximating the confidence intervals for Sharpe style weights“. In: *Financial Analysts Journal* 53.4, pp. 80–85.

Löfberg, J. (2004). „YALMIP : A Toolbox for Modeling and Optimization in MATLAB“. In: *In Proceedings of the CACSD Conference*. Taipei, Taiwan.

Ma, Lingjie and Larry Pohlman (2008). „Return forecasts and optimal portfolio construction: a quantile regression approach“. In: *The European Journal of Finance* 14.5, pp. 409–425.

MacQueen, Jason (2011). *Portfolio Risk Decomposition (and Risk Budgeting)*. Tech. rep. R-Squard Risk Management Research.

Mahalanobis, Prasanta Chandra (1936). „On the generalized distance in statistics“. In: *Proceedings of the National Institute of Sciences of India* 2.1, pp. 49–55.

Malz, Allan M (2014). *A Simple and Reliable Way to Compute Option-based Risk-neutral Distributions*. Tech. rep. 677. FRB of New York.

Mandelbrot, Benoit B (2013). *Fractals and scaling in finance: Discontinuity, concentration, risk. Selecta volume E*. Springer Science & Business Media.

Mandelbrot, Benoit B and John W Van Ness (1968). „Fractional Brownian motions, fractional noises and applications“. In: *SIAM review* 10.4, pp. 422–437.

Markowitz, Harry (1952). „Portfolio selection“. In: *The journal of finance* 7.1, pp. 77–91.

Martin, Ian and Stephen Ross (2013). *The Long Bond*. Tech. rep. London School of Economics and MIT.

McManus, Desmond John (1999). „The information content of interest rate futures options“. In: *Working Paper*, pp. 99–115.

Melick, William R and Charles P Thomas (1997). „Recovering an Asset's Implied PDF from Option Prices: An Application to Crude Oil during the Gulf Crisis“. In: *Journal of Financial and Quantitative Analysis* 32.01, pp. 91–115.

Merton, Robert C (1973a). „An intertemporal capital asset pricing model“. In: *Econometrica* 41.5, pp. 867–887.

– (1973b). „Theory of Rational Option Pricing“. In: *The Bell Journal of Economics and Management Science*, pp. 141–183.

– (1976). „Option Pricing when Underlying Stock Returns are Discontinuous“. In: *Journal of Financial Economics* 3.1-2, pp. 125–144.

Meucci, Attilio (2001). „Common pitfalls in mean-variance asset allocation“. In: *Wilmott Magazine*.

– (2007). „Risk contributions from generic user-defined factors“. In: *The Risk Magazine*, pp. 84–88.

– (2009). „Managing Diversification“. In: *Risk* 22.5, pp. 74–79.

Michaud, Richard O (1989). „The Markowitz optimization enigma: Is 'optimized' optimal?“ In: *Financial Analysts Journal* 45.1, pp. 31–42.

Mohapi, Tjhaka and Ilsa Botha (2013). „The Explanatory Power Of The Yield Curve In Predicting Recessions In South Africa“. In: *International Business & Economics Research Journal (IBER)* 12.6, pp. 613–634.

Moolman, Elna (2003). „Predicting turning points in the South African economy“. English. In: *South African Journal of Economic and Management Sciences* 6.2, pp. 289–303.

Morris, Quinton, Gary Van Vuuren, and Paul Styger (2009). „Further evidence of long memory in the South African stock market“. In: *South African Journal of Economics* 77.1, pp. 81–101.

Moskowitz, Tobias J, Yao Hua Ooi, and Lasse Heje Pedersen (2012). „Time series momentum“. In: *Journal of financial economics* 104.2, pp. 228–250.

Muller, Chris and Mike Ward (2013). „Style-based effects on the Johannesburg Stock Exchange: A graphical time-series approach“. In: *Investment Analysts Journal* 42.77, pp. 1–16.

Müller, Ulrich A, Michel M Dacorogna, and Oliver V Pictet (1998). „Heavy tails in high-frequency financial data“. In: *A practical guide to heavy tails: Statistical techniques and applications*, pp. 55–78.

Mutooni, Rand and C Muller (2007). „Equity style timing“. In: *Investment Analysts Journal* 36.65, pp. 15–24.

Mutswari, Phathutshedzo (2016). „On the hunt for a factor model of South African stock returns“. MPhil Dissertation. University of Cape Town.

Necula, Ciprian (2002). „Option Pricing in a Fractional Brownian Motion Environment“. In: *SSRN Electronic Journal* 1286833.

Novy-Marx, Robert (2013). „The other side of value: The gross profitability premium“. In: *Journal of Financial Economics* 108.1, pp. 1–28.

Nystrup, Peter (2014). „Regime-Based Asset Allocation“. MA thesis. PhD thesis, Technical University of Denmark.

Park, Sungyong, Wendun Wang, and Naijing Huang (2015). *Testing for Stock Market Contagion: A Quantile Regression Approach*. Tech. rep. Tinbergen Institute.

Peters, Edgar E (1989). „Fractal structure in the capital markets“. In: *Financial Analysts Journal* 45.4, pp. 32–37.

Peters, Edgar E, Edgar R Peters, and Donada Peters (1994). *Fractal market analysis: applying chaos theory to investment and economics*. Vol. 24. John Wiley & Sons.

Podkaminer, E (2013). „Risk factors as building blocks for portfolio diversification: The chemistry of asset allocation“. In: *CFA Institute Investment Risk and Performance* 2013.1, pp. 1–15.

Posner, Steven E and Moshe A Milevsky (1998). „Valuing Exotic Options by Approximating the SPD with Higher Moments“. In: *The Journal of Financial Engineering* 7.2.

Qin, Likuan and Vadim Linetsky (2016). „Positive eigenfunctions of markovian pricing operators: Hansen-scheinkman factorization, ross recovery, and long-term pricing“. In: *Operations Research* 64.1, pp. 99–117.

Rejichi, Imen Zgueb and Chaker Aloui (2012). „Hurst exponent behavior and assessment of the MENA stock markets efficiency“. In: *Research in International Business and Finance* 26.3, pp. 353–370.

Richard, Jean-Charles and Thierry Roncalli (2015). „Smart Beta: Managing Diversification of Minimum Variance Portfolios“. In: *Risk-Based and Factor Investing*. Ed. by Emmanuel Jurczenko. Elsevier, pp. 31 –63.

Rogers, Leonard CG and MR Tehranchi (2010). „Can the implied volatility surface move by parallel shifts?“ In: *Finance and Stochastics* 14.2, pp. 235–248.

Roncalli, Thierry (2013). *Introduction to risk parity and budgeting*. CRC Press.

Roncalli, Thierry and Guillaume Weisang (2016). „Risk parity portfolios with risk factors“. In: *Quantitative Finance* 16.3, pp. 377–388.

Roper, Michael (2010). *Arbitrage Free Implied Volatility Surfaces*. Tech. rep. University of Sydney Working Paper.

Ross, Stephen (2015). „The Recovery Theorem“. In: *The Journal of Finance* 70.2, pp. 615–648.

Ross, Stephen A (1976). „The Arbitrage Theory of Capital Asset Pricing“. In: *Journal of Economic Theory* 13.3, pp. 341–360.

Russo, Alessandro (2015). *Equity Factor investing According to the Macroeconomic Environment.* Tech. rep. Amundi Asset Management.

Sapra, S and M Hunjan (2013). *Active Share, Tracking Error and Manager Style*. Tech. rep. PIMCO.

Schlanger, Todd, Christopher Philips, and Karin Peterson LaBarge (2012). *The Search for Outperformance: Evaluating 'Active Share'*. Tech. rep. Vanguard Research.

Serinaldi, Francesco (2010). „Use and misuse of some Hurst parameter estimators applied to stationary and non-stationary financial time series“. In: *Physica A: Statistical Mechanics and its Applications* 389.14, pp. 2770–2781.

Seymour, Anthony (2011). *Exotic Options for Fund Managers: Analytic Contributions.* Tech. rep. Legae Peresec Research.

Seymour, Anthony, Florence Chikurunhe, and Emlyn Flint (2012). „Hedge Design: A Systematic Approach“. In: *SSRN Electronic Journal* 2767414.

Seymour, Anthony, Emlyn Flint, and Florence Chikurunhe (2014). „Single Stock and Custom Basket Options in Fund Management“. In: *SSRN Electronic Journal* 2767424.

Seymour, Anthony, Florence Chikurunhe, and Emlyn Flint (2015). „Currency Hedge Design: Accounting for Uncertain Correlation and Volatility“. In: *SSRN Electronic Journal* 2767437.

Seymour, Anthony, Emlyn Flint, and Florence Chikurunhe (2016). *Peregrine Regime Market Utility*. Tech. rep. Legae Peresec Research.

Sharpe, William F. (1964). „Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk“. In: *The Journal of Finance* 19.3, pp. 425–442.

– (1992). „Asset allocation: Management style and performance measurement“. In: *The Journal of Portfolio Management* 18.2, pp. 7–19. eprint: `https://jpm.iijournals.com/content/18/2/7.full.pdf`.

Shimko, David (1993). „The Bounds of Probability“. In: *RISK* 6, pp. 33–37.

Shreve, Steven E (2004). *Stochastic Calculus for Finance II: Continuous-time Models*. Vol. 11. Springer Science & Business Media.

Sim, Nicholas (2016). „Modeling the dependence structures of financial assets through the Copula Quantile-on-Quantile approach“. In: *International Review of Financial Analysis* 48, pp. 31–45.

Simonsen, Ingve (2003). „Measuring anti-correlations in the Nordic electricity spot market by wavelets“. In: *Physica A: Statistical Mechanics and its applications* 322, pp. 597–606.

Spears, Trent (2013). „On Estimating the Risk-Neutral and Real-World Probability Measures“. MA thesis. Oxford University.

Strugnell, Dave, Evan Gilbert, and Ryan Kruger (2011). „Beta, size and value effects on the JSE, 1994-2007“. In: *Investment Analysts Journal* 40.74, pp. 1–17.

Tikhonov, AN and V Ya Arsenin (1977). *Methods for Solving Ill-Posed Problems*. John Wiley and Sons, Inc.

Tompkins, Robert G (2001). „Implied Volatility Surfaces: Uncovering Regularities for Options on Financial Futures“. In: *The European Journal of Finance* 7.3, pp. 198–230.

Tran, Ngoc-Khanh and Shixiang Xia (2015). *Specified Recovery*. Tech. rep. Washington University in St. Louis.

Treynor, Jack and Kay Mazuy (1966). „Can mutual funds outguess the market“. In: *Harvard business review* 44.4, pp. 131–136.

Tzouras, Spilios, Christoforos Anagnostopoulos, and Emma McCoy (2015). „Financial time series modeling using the Hurst exponent“. In: *Physica A: Statistical Mechanics and its Applications* 425, pp. 50–68.

Van Rensburg, Paul (2001). „A decomposition of style-based risk on the JSE“. In: *Investment Analysts Journal* 30.54, pp. 45–60.

Van Rensburg, Paul and Michael Robertson (2003). „Size, price-to-earnings and beta on the JSE Securities Exchange“. In: *Investment Analysts Journal* 32.58, pp. 7–16. eprint: `https://doi.org/10.1080/10293523.2003.11082449`.

Van Vliet, Pim and David Blitz (2011). „Dynamic strategic asset allocation: Risk and return across the business cycle“. In: *Journal of Asset Management* 12.

Venkatramanan, Aanand and Carol Alexander (2011). „Closed Form Approximations for Spread Options“. In: *Applied Mathematical Finance* 18, pp. 1–26.

Walden, Johan (2014). „Recovery with Unbounded Diffusion Processes“. In: *SSRN Electronic Journal* 2508414.

West, Graeme (2005). „Calibration of the SABR Model in Illiquid Markets“. In: *Applied Mathematical Finance* 12.4, pp. 371–385.

Xu, Fengmin, Zhaosong Lu, and Zongben Xu (2016). „An efficient optimization approach for a cardinality-constrained index tracking problem“. In: *Optimization Methods and Software* 31.2, pp. 258–271.

Ye, Wuyi, Yangguang Zhu, Yuehua Wu, and Baiqi Miao (2016). „Markov regime-switching quantile regression models and financial contagion detection“. In: *Insurance: Mathematics and Economics* 67, pp. 21–26.

Zucchini, Walter, Iain L MacDonald, and Roland Langrock (2016). *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC.

Websites

Meucci, Attilio (2019). *Advanced Risk and Portfolio Management Lab*. ARPM. URL: https://www.arpm.co/lab/.

Varadi, David (2014). *Are Simple Momentum Strategies Too Dumb? Introducing Probabilistic Momentum*. CSSA Analytics. URL: https://cssanalytics.wordpress.com/2014/01/28/are-simple-momentum-strategies-too-dumb-introducing-probabilistic-momentum/.

A

Calibration Algorithms and Regularisation Parameters

A.1 Calibration Algorithm for the SVI Implied Volatility Model

We summarise the SVI calibration procedure given in De Marco and Martini (2009). Define $w(x) = \tau\sigma^2(x)$ as a single total variance skew and introduce a change of variables

$$y = \frac{x - m}{s}. \tag{A.1}$$

Using this new variable we can rewrite the SVI equation in total variance space as

$$w(y) = \alpha + \delta y + \beta\sqrt{y^2 - 1}, \tag{A.2}$$

where

$$\begin{aligned} \alpha &= a\tau \\ \beta &= bs\tau \\ \delta &= \rho bs\tau. \end{aligned} \tag{A.3}$$

Taking m and s as fixed values, the objective function then becomes

$$\underset{(\alpha,\beta,\delta)\in D}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(\alpha + \delta y + \beta\sqrt{y^2 - 1} - V_i\right)^2, \tag{A.4}$$

where n is the total number of market observations for the given expiry and $V_i = v_i\tau$ is the i^{th} observed total variance. The domain D is defined as

$$D = \begin{cases} 0 \leq \beta \leq 4s \\ |\delta| \leq \beta \text{ and } |\delta| \leq 4s - \beta \\ 0 \leq \alpha \leq max\{\nu_i\}. \end{cases} \tag{A.5}$$

For a given solution set $\{\alpha^*, \beta^*, \delta^*\}$, one can find the corresponding solution set $\{a^*, b^*, \rho^*\}$ and thus the remaining parameters can be calibrated as follows

$$\underset{m,s\geq 0}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(\sigma^2(x, \{m, s, a^*, b^*, \rho^*\}) - v_i\right)^2. \tag{A.6}$$

The original five-dimensional calibration is thus broken into separate three-dimensional and two-dimensional minimisation problems.

A.2 Regularisation Parameters for the Recovery Theorem

Kiriu and Hibiki (2015) calculate the regularisation target matrix $\bar{P}$ directly from the discretised RND matrix Q based on two premises. Firstly, because the expiry in the first column of Q is equal to the expiry of the transition probability matrix P, and because the states are chosen symmetrically around the current market level, it must be that the middle row of the P matrix is equal to the first column of Q. Secondly, Kiriu and Hibiki (2015) suggest that the probability of transitioning from states S_i to S_j should be similar to the probability of transitioning from states S_{i+k} to S_{j+k} for all $k \leq min\left(n-i, n-j\right)$. The first premise defines the middle row of $\bar{P}$ while the second defines the remainder of the matrix. Assuming that one has an odd number n of states and defining $m_0 = (n+1)/2$ as the middle state, we have that

$$\bar{P} = \begin{bmatrix} \sum_{i=1}^{m_0} q_{i,1} & q_{m_0+1,1} & \cdots & q_{n,1} & \cdots & 0 & 0 \\ \sum_{i=1}^{m_0-1} q_{i,1} & q_{m_0,1} & \cdots & q_{n-1,1} & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots \\ q_{1,1} & q_{2,1} & \cdots & q_{m_0,1} & \cdots & q_{n-1,1} & q_{n,1} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & q_{2,1} & \cdots & q_{m_0,1} & \sum_{i=m_0+1}^{n} q_{i,1} \\ 0 & 0 & \cdots & q_{1,1} & \cdots & q_{m_0-1,1} & \sum_{i=m_0}^{n} q_{i,1} \end{bmatrix}. \tag{A.7}$$

Although there a number of standard functions used to evaluate regularisation parameters, Kiriu and Hibiki (2015) suggest a problem-specific selection function that attempts to balance relative gain in the objective function from each term in the regularised OLS minimisation:

$$\begin{aligned} P &= \underset{p_{ij} \geq 0}{\text{argmin}} \left\| AP - B \right\|_2^2 + \zeta \left\| P - \bar{P} \right\|_2^2 \\ &= \underset{p_{ij} \geq 0}{\text{argmin}}\, y_{fit}\left(\zeta\right) + \zeta y_{reg}\left(\zeta\right). \end{aligned} \tag{A.8}$$

The selection function $h(\zeta)$ is then given as

$$h\left(\zeta\right) = \frac{y_{fit}\left(\zeta\right) - y_{fit}\left(0\right)}{y_{fit}\left(\infty\right) - y_{fit}\left(0\right)} + \frac{y_{reg}\left(\zeta\right) - y_{reg}\left(\infty\right)}{y_{reg}\left(0\right) - y_{reg}\left(\infty\right)}, \tag{A.9}$$

where the respective denominators represent the maximum spread in each term and the numerator gives the spread achieved for a specified ζ value. The $y_i\left(0\right)$ values

are solutions from the original OLS problem and $y_{reg}(\infty)$ is set to 0 due to the fact that $P \to \bar{P}$ as $\zeta \to \infty$. This implies that $h(0) = h(\infty) = 1$. Kiriu and Hibiki (2015) show under simulation that the h function is smooth, continuous and has a single minimum value. Most importantly, the derivative function h' is very stable around this global minimum, making it a very appealing selection function.

Colophon

This thesis was typeset with LaTeX 2$_\varepsilon$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Declaration

I, the undersigned, declare that this thesis, which I hereby submit for the degree *Philosophiae Doctor* at the University of Pretoria, is my own independent work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Pretoria, 14 July 2019

Emlyn Flint