Cross asset class applications of functional data analysis: Evaluation with controls for data snooping bias

Fearghal J. Kearney B.Sc.

Internal supervisor: Dr Mark Cummins, Dublin City University Business School

External supervisor: Dr Finbarr Murphy, Kemmy Business School, University of Limerick

December 2014

A thesis submitted to Dublin City University Business School in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____

ID No.: 56372201

Date:

Acknowledgments

I am indebted to my supervisors, Dr Mark Cummins and Dr Finbarr Murphy. Mark and Finbarr have provided great support and guidance during the past three years, both in relation to this thesis and my broader academic training. They have helped to dramatically improve my writing and analytical skills. It has been a pleasure working with them both.

The financial support of DCU Business School and the Irish Accounting and Finance Association is gratefully acknowledged. I would like to thank Professor Liam Gallagher, Professor Ronan Powell, Professor Brian O'Kelly, Dr Hiroyuki Kawakstu, Dr Michael Dowling, Mr Billy Kelly, Professor Ana-Maria Fuertes, Professor William Bertin, Mr Daire McCoy, and Dr Surajit Ray for their comments, advice, and guidance during my studies. I reserve special thanks for Dr Teresa Hogan who consistently went beyond the call of duty in her former role as Director of Doctoral Studies.

I would like to thank my family, Paddy, Eileen, Cormac, Dearbhla, and Donncha for their emotional and financial support. I am eternally grateful to my late brother Cian, who I miss each and every day, for the many nights he pulled up a chair, characteristically exuding patience and logic, just as the search for the correct integration substitution turned from challenge to frustration. Finally, I thank Davina for her warmth, friendship and love.

Contents

List of Figures

List of Tables

Cross asset class applications of functional data analysis: Evaluation with controls for data snooping bias

Fearghal J. Kearney

Abstract

This thesis applies functional data analysis techniques to address a number of specific research questions in financial markets. Data snooping bias controls are adopted in parallel to provide statistical robustness to our inferences. Firstly, we conduct an investigation into U.S. exchange-traded fund outperformance during the 2008-2012 period. The funds are tested for net asset value premium, underlying index and market benchmark outperformance. The study serves as a platform to showcase the data snooping bias problem and application of generalised multiple hypothesis testing techniques, in advance of their use for functional data analysis evaluation. Secondly, as the first application of functional data analysis, we examine implied volatility, jump risk, and pricing dynamics within crude oil markets. Strong evidence is found of converse jump dynamics during periods of demand and supply side weakness. Next, we demonstrate the performance advantage over traditional benchmarks of adopting a functional linear model to forecast EUR-USD implied volatility. Our findings are shown to be robust across various moneyness segments, contract maturities and out-of-sample window lengths. The final chapter also uses a functional data framework to produce forecasts, demonstrating how information can be extracted from forward contracts to predict future spot foreign exchange rates. The evaluation of an out-of-sample framework leads to near systematic outperformance in terms of a direct comparison of performance measures, versus both the restricted vector error correction model and random walk. Overall, this thesis highlights the usefulness of adopting insightful and novel functional data analysis techniques across various asset classes where multiple hypothesis testing controls provide robustness around our conclusions. Each of the studies contributes to the literature individually, with the collection emphasising the benefits of adopting functional approaches to tackle a wide range of empirical finance problems.

Chapter 1

Introduction

1.1 Introduction

The core contribution of the thesis is the proposed use of functional data analysis (FDA) techniques in a cross-security financial setting. Multiple hypothesis testing (MHT) controls are adopted in parallel to robustly identify instances of outperformance. We demonstrate the power and flexibility of the statistical techniques by applying them across four distinct asset classes, namely, exchange-traded funds (ETFs), crude oil options, foreign exchange implied volatility, and foreign exchange forward rate term structure. We outline the use of recent innovations in controlling for the MHT problem when seeking to identify ETF outperformance and use functional data analysis techniques to examine implied volatility, jump risk, and pricing dynamics within crude oil markets. Furthermore, we combine both FDA and MHT techniques to characterise and forecast underlying processes; forecasting EUR-USD implied volatility and extracting the informational content from the forward rate term structure for multiple currencies.

1.2 Context and motivation

Functional data analysis (FDA) provides a framework to produce and interpret functional representations of the process underlying a data set. Functional data analysis begins with the assumption that there exists an underlying function that generates the observations. In addition, it is assumed that the underlying function is smooth in some sense, so that there is a link between consecutive observations. The process is defined over a continuum, where continuum values are generally represented in terms of time or space. This continuous property distinguishes FDA from other common multivariate techniques which seek to model and forecast financial processes based solely on the discrete observations observed in the data set. In this thesis the functions are defined over the domain spanned by both the moneyness and deltas of option contracts, and the tenors of forward contracts. The resultant functions serve to characterise the implied volatility curve and forward rate term structure.

Functional data analysis is adopted for this thesis as it boasts many advantages; it accurately captures underlying smooth process dynamics, there is no assumed parametric structure, it is computationally efficient, and it results in a representation that can be evaluated on an arbitrarily fine grid (Ramsay and Silverman 2005). Another significant feature that we exploit to obtain a proxy for jump risk, is that the continuous function can be differentiated at any point to obtain the slope and other higher order derivatives. Functional variants of many traditional multivariate techniques are available with a number of applications in the bio-mechanical literature benefitting from functional data frameworks. However, it has only recently been exploited for financial analysis with studies by Benko et al. (2009) and Muller et al. (2011) highlighting its usefulness in characterising implied and realised volatility processes, respectively. We adopt a distinct functional framework to also forecast implied volatility, and further employ functional techniques to characterise the forward rate term structure in foreign exchange markets. The framework can be used as an exploratory tool to represent and analyse a financial data set in infinite dimensional space, such as conducted for crude oil options but also to predict the future evolution of financial market processes.

When constructing a functional data object, a vector of n bases, denoted $\phi_1, ..., \phi_n$, must first be specified. The decision of which basis system to specify is driven by the underlying data's known characteristics. For instance, when modelling periodic data, a Fourier basis expansion, comprised of successive sine/cosine terms, is most commonly applied. However, neither the implied volatility or forward curves exhibit strong cyclical variation, so we choose flexible B-splines for the basis function system. B-spline representation offers a number of strengths, as outlined in de Boor (2001). They are essentially a number of polynomials joined together smoothly at fixed points called knots. The number and positioning of the knots are derived from knowledge of the complexity of the underlying process over particular ranges. Computations with B-splines are extremely efficient as at any one point along the curve they simplify to a polynomial that can be easily evaluated. Adjusting the order of the spline allows for the estimation of derivatives of any degree. Functional structures are approximated as a weighted linear combination of these bases. We employ a number of distinct functional regression specifications to forecast these structures. Classical linear models seek to describe the dependency between a response variable and a specified set of predictors. In classical regression, scalar values are used for both the explanatory and response variables. However, in functional linear regression at least one of the observed variables is a curve.

The motivation for incorporating multiple hypothesis testing controls is to account for the issue of data snooping bias. This ensure that the conclusions drawn around the use of our functional data frameworks in the applied studies are as robust as possible. Data snooping bias, in this context, is the problem whereby under näive analysis statistically significant outperformance relationships may be identified by pure chance alone. When conducting a number of hypothesis tests simultaneously on the same data set one runs the risk of uncovering random artefacts as statistically significant relationships. This is

due to inherent correlations observed between members of any "random" data set. The false discovery of such random artefacts can inhibit risk management and the pricing and hedging of derivatives. Data snooping bias links directly to the broader issue of multiple hypothesis testing in statistical and econometric applications where the issue is commonly referred to as the multiple comparisons problem. Data snooping bias is well addressed in scientific and medical fields but largely ignored in empirical finance literature. A number of quantitative studies employ such procedures. Sullivan and Timmermann (1999), Hsu and Kuan (2005), Park and Irwin (2007), Marshall et al. (2008), and Qui and Wu (2008) apply the reality check bootstrap test of White (2000) to evaluate the profitability of a wide range of technical trading rules commonly used in industry. Further to this, Hsu et al. (2010) employ a stepwise extension of the superior predictability test of Hansen (2005) to re-evaluate the profitability of technical trading rules. The methodologies used in previous studies raise concerns around the validity of the inferences drawn, insofar as they lack data snooping bias controls and in many cases conduct less formal hypothesis tests. This can greatly mislead an investor's portfolio selection. Addressing this issue is important as it calls into question, and potentially undermines, the findings and conclusions in the literature. A major contribution of the thesis to the literature is the utilisation of a generalised data snooping bias procedure in the performance appraisal setting. We apply the generalised balanced stepdown procedure of Romano and Wolf (2010), which serves as an improvement over the more conservative seminal reality check bootstrap test of White (2000) and the superior predictive ability test of Hansen (2005). The generalised balanced stepdown procedure of Romano and Wolf (2010) boasts a greater ability to reject false null hypotheses as well as offering balance in the sense that all hypotheses are treated equally in terms of power. It also allows for subsequent iterative steps to identify additional hypothesis rejections. The technique is outlined in Chapter 2, and is also applied in Chapters 4 and 5.

1.3 Contribution

1.3.1 Research questions

We will now outline the reasons for our choice of asset class concentrations and research questions. Exchange-traded funds (ETFs) are variants of mutual funds that first came to prominence in the early 1990s. ETFs allow market participants to trade index portfolios, similar to how individual investors *trade* shares of a stock. They seek to track the value and volatility of an underlying benchmark index through the construction of portfolios replicative of the index's constituents. They were first traded on the Toronto Stock Exchange in 1989 and today's market boasts over 1,220 U.S. traded ETFs.[1] In relation to the analysis of ETF outperformance, the majority of research conducted to date has centered on data sets comprising small numbers of large ETFs, single ETF families or industries,

[1] Investment company institute June 2012 ETF report:
http://www.ici.org/etf_resources/research/etfs_06_12.

with measurements being applied inconsistently across the differing studies, inhibiting effective cross comparison. This thesis amends that, primarily through the use of a large, diverse sample size, which incorporates many sectoral and internationally focused indices.

We investigate a large number of ETF attributes and their ability to dictate net asset value premium, underlying index and market benchmark outperformance. The effect of replication type and asset class focus on ETF performance for instance has not been rigorously tested in the literature and as such we incrementally contribute in this way. This work may be of interest to a variety of stakeholders. Firstly, investigating ETF outperformance is significant from an academic perspective as it furthers our understanding of the market's pricing dynamics. Secondly, the wider investment community would benefit from the work as an aid in identifying specific ETF cohorts suitable to individual portfolio requirements. Lastly, despite it not being the primary focus of this study, the data snooping bias issues raised offer broader insights to arbitrageurs by emphasising the importance of controlling for data snooping bias in order to robustly identify mispricings and trading signals.

Oil futures are the most actively traded commodity derivatives. An average of one million light sweet crude oil futures and option contracts are traded every day according to the CME group.[2] The past 10 years have seen elevated levels of price volatility in these markets. Strong economic pressures have been observed on both the demand side and the supply side, during the global financial crisis and the Arab Spring respectively. Increased price volatility in oil markets causes profound economic management and socio-political issues, not only impacting those participants who invest directly in commodities but also the consumers of refined oil products. There is a large body of literature demonstrating the importance of incorporating jumps into models seeking to capture risk premia and economic shocks. Traditional geometric Brownian motion based models, such as Black and Scholes' (1973) diffusion model, do not capture price jumps, which are movements that become more prevalent during periods of increased market turbulence. For this reason we employ the use of the Merton (1976) model in line with Yan (2011). Yan (2011) proposes the use of implied volatility slope information to estimate jump risk. He shows, both directly and indirectly, the applicability of the at-the-money implied volatility slope as a proxy for the average jump amplitude in equity markets. We seek to answer a similar question in crude oil markets. Further contributions relate to the employment of FDA-obtained Merton model parameters for portfolio hedging where we compare the calculated results with the standard Black-Scholes delta hedging strategy.

Observed implied volatility differs across option contracts, dependent on both moneyness and expiry date. As well as being a transformation of the option price, and a key parameter in many asset pricing formulae, implied volatility is also of interest due to its informational content (see Corrado and Miller 2006, Taylor et al. 2010, Muzzioli 2010, and Garvey and Gallagher 2012). Yu et al. (2010) demonstrate this by finding superior results using implied volatility to predict future return volatility of stock index options,

[2] http://www.cmegroup.com/trading/energy/files/en-153_wti_brochure_sr.pdf.

when compared to traditional benchmark models in over-the-counter (OTC) and exchange markets. One such OTC market is that of foreign exchange (FX) options. FX is the largest asset class in the world with the Bank for International Settlements reporting that trading levels in FX markets averaged $5.3 trillion per day.[3] Many stakeholders are exposed to FX risk including banks, speculators, traders, multinational firms, importers, and exporters. Modelling foreign currency cash flows, investment decisions, and hedging strategies, are all greatly dependent on expectations of future FX movements. Relative to previous studies forecasting the volatility of returns, there is a relative paucity of literature predicting the evolution of implied volatility.

We add to the existing FX implied volatility literature through the novel proposal of a functional data analysis-based forecasting model to predict the evolution of the implied volatility function. The aim is to determine and forecast the function that characterises the implied volatility relationship among option contracts. Both the scalar response/functional explanatory and functional explanatory/functional response linear models of Ramsay and Silverman (2005) are utilised for the analysis, with the forecasts compared to traditionally proposed benchmarks of Gonclaves and Guidolin (2006) and Konstantinidi et al. (2008), in an out-of-sample testing framework. We not only contribute from an academic perspective, where insights into the dynamics of implied volatility aid our understanding of option markets, but also from a market practitioner perspective, due to the study's potential hedging and speculation implications. We contribute further by incorporating the use of a contributory data vendor. This mitigates the idiosyncratic risk, as highlighted by Chalamandaris and Tsekrekos (2014), associated with obtaining quotes from a single market participant.

Meese and Rogoff (1983a,b) ascertain that standard exchange rate models do not have the ability to beat forecasts implied by the random walk in the short run. In an attempt to explain this, Engel and West (2005) and Engel et al. (2008) demonstrate that such models imply a near random walk process for the exchange rate, so their power to "beat the random walk" in out-of-sample forecasts is low. Furthermore, it has been demonstrated that the forward rate is not the optimal predictor of future spot rates (Hansen and Hodrick 1980, Frankel 1980, Bilson 1981, Frankel and Rose 1995, and Taylor 1995). Despite this, the question as to whether or not there is information imbedded in forward FX rates persists. Clarida and Taylor (1997) seek to answer this by moving beyond such single-equation methods and conclude that forward premia information is in fact considerable. Their restricted vector error correction model (VECM) constitutes the leading challenger to the seminal work of Meese and Rogoff (1983a,b). The approach is applied in a dynamic out-of-sample forecasting framework resulting in root mean squared error and mean absolute error metrics over 50% lower than those implied by the random walk. The results are confirmed by Clarida et al. (2003) and Sager and Taylor (2014), who establish statistically significant outperformance across different data sets. Our study adds to the existing literature seeking to extract the informational content of forward foreign exchange rates through the novel

[3] http://www.bis.org/publ/rpfx13fx.pdf.

proposal of a functional data analysis-based forecasting model.

In relation to extracting the informational content of the FX forward curve, we contribute by moving the problem to a functional space to improve on the forecasting performance achieved by previous leading benchmark models. To this aim, we adopt the scalar response model proposed in Ramsay and Silverman (2005) where the infinite dimensional beta coefficient is specified with a functional principal component basis to solve the underdetermination issue. Specifically, we determine the underlying process that characterises the forward rate term structure to establish dependency relations between forward rates and future spot exchange rates. The flexible functional data approach accurately captures the forward rate term structure process, whilst mitigating the need to impose restrictive data structure assumptions on the exchange rate system. For comparative purposes with previous studies, we initially present a direct comparison of forecasting performance measures. However, we then apply formal tests to identify instances of statistically significant outperformance for the scalar response model over both the VECM and random walk benchmarks. We first test the hypothesis of forecasting outperformance by implementing a simple t-test of performance measures differentials. In an important extension of the literature we contribute by incorporating controls for the multiple comparisons problem in testing forecast performance. Further to this, our framework tests if exchange rates are in fact predictable and if the simple risk neutral efficient market hypothesis holds.

1.3.2 Chapter outline

Chapter 2, "Outperformance in exchange-traded fund pricing deviations: Generalised control of data snooping bias", conducts an investigation into exchange-traded fund (ETF) outperformance during the 2008-2012 period, utilising a data set of 288 U.S. traded securities. ETFs are tested for net asset value (NAV) premium, underlying index and market benchmark outperformance, with Sharpe, Treynor, and Sortino ratios employed as risk-adjusted performance measures. A key contribution is the application of an innovative generalised stepdown procedure in controlling for data snooping bias.

Chapter 2 key questions:

- What ETFs display a net asset value premium?
- What ETFs outperform their underlying index?
- What ETFs outperform market wide benchmarks?
- How do specific groupings of ETFs differ in terms of outperformance?

Chapter 2 key findings:

- Energy, Precious Metals, Real Estate and Leisure industries beat the market on a risk adjusted basis.
- Powershares DB Silver and iShares Silver Trust substantially outperform.
- 63% and 79% of Global and International ETFs respectively, show premium Sharpe Ratio outperformance with only 10% for US funds.
- ETFs exhibiting high expense ratios or recent inception dates have a greater tendency to outperform their index.

Chapter 3, "An analysis of implied volatility jump dynamics: Novel functional data representation in crude oil markets", proposes a framework to produce and interpret functional objects that characterise the underlying dynamics of oil future options. The functional data analysis framework is used to examine implied volatility, jump risk, and pricing dynamics within crude oil markets. Examining a WTI crude oil sample for the 2007-2013 period, which includes the global financial crisis and the Arab Spring, strong evidence is found of converse jump dynamics during periods of demand and supply side weakness. This is used as a basis for an FDA-derived Merton (1976) jump diffusion optimised delta hedging strategy, which exhibits superior portfolio management results over traditional methods.

Chapter 3 key questions:

- What is the link between the shape of the implied volatility smile and underlying economic events in crude oil markets?
- Does the implied volatility curve slope contain information useful in specifying the average jump amplitude for crude oil options, in a similar manner to what Yan (2011) has shown to be the case for stock returns?
- Can information contained in the implied volatility smile slope be exploited to improve portfolio hedging techniques?

Chapter 3 key findings:

- Strong evidence is found of converse jump dynamics during periods of demand and supply side weakness
- An FDA-derived Merton (1976) jump diffusion optimised delta hedging strategy exhibits superior portfolio management results over traditional methods

Chapter 4, "Forecasting implied volatility in foreign exchange markets: A robust functional linear model approach", utilises functional data analysis techniques to characterise and forecast implied volatility in foreign exchange markets. The process of interest in this study is that of the EUR-USD daily implied volatility curve. Superior prediction of the evolution

of the implied volatility process is exhibited. This evaluation is performed under a rigorous out-of-sample testing framework that controls for the multiple comparisons problem.

Chapter 4 key questions:

- Can functional linear model techniques be used to characterise and forecast implied volatility in foreign exchange markets?
- How does the performance of the functional data analysis approach compare to traditionally employed benchmark models of Gonclaves and Guidolin (2006) and Konstantinidi et al. (2008)?
- Are the findings robust across various moneyness segments, contract maturities and out-of-sample window lengths?

Chapter 4 key findings:

- Our FDA techniques uncover predictable patterns in implied volatility
- We robustly demonstrate the performance advantage of adopting an FDA framework when predicting future implied volatility
- We empirically demonstrate that the specification of a scalar response model provides a superior implied volatility fit over the fully functional model.

Chapter 5, "Extracting FX forward rate term structure information: Merits of a functional method", seeks to extract the informational content of the forward rate term structure through the implementation of a functional principal component-based scalar response model. The difficulty of beating the random walk in forecasting spot FX rates is well documented, with the restricted VECM of Clarida and Taylor (1997) providing the primary challenge. Our out-of-sample framework leads to near systematic outperformance in terms of a direct comparison of performance measures, versus both the VECM and random walk.

Chapter 5 key questions:

- Can we extract the informational content of forward foreign exchange rates through a functional PCA-based forecasting model?
- How does the performance of the functional PCA-based approach compare with both the random walk and the Clarida and Taylor (1997) VECM?
- Does the forward rate term structure contain information about the evolution of spot exchange rates?

Chapter 5 key findings:

- Our scalar response model leads to near systematic outperformance in terms of a direct comparison of performance measures, coupled with multiple instances of truly significant outperformance versus both the random walk and Clarida and Taylor (1997) VECM
- Our results indicate that the forward rate term structure contains statistically significant information about the evolution of the spot exchange rate
- We provide additional evidence supporting the rejection of the simple risk neutral efficient market hypothesis

Chapter 6 outlines the major conclusions drawn from the work.

1.4 Research dissemination

This thesis resulted in a paper that was published in the *Journal of Financial Markets*; Kearney, F., M. Cummins, and F. Murphy. 2014. Outperformance in exchange-traded fund pricing deviations: Generalized control of data snooping bias. *Journal of Financial Markets* 19:86-109. A manuscript based on Chapter 3 received a revise and resubmit decision from *The North American Journal of Economics and Finance* (ranked 16th/89 in the Business, Finance category of the Thomson Reuters ISI list). It has since been amended, in line with the version presented here and will be resubmitted in the near future. Chapters 4 and 5, "Forecasting implied volatility in foreign exchange markets: A robust functional linear model approach" and "Extracting FX Forward Rate Term Structure Information: Merits of a Functional Method", are both working papers and will be submitted soon. Research from this thesis have been presented at the Infiniti Conference, Prato 2014, the Infiniti Conference, Aix-en-Provence 2013, the Irish Society of New Economists (ISNE) Conference, UCC 2012, the ISNE Conference, NUIM 2013, the ISNE Conference, NUIG 2014, the DCU brown bag seminar, and the DCU doctoral colloquium. This research also resulted in the award of the DCU Business School scholarship and separately, in the receipt of the Irish Accounting and Finance Association research funding bursary.

Chapter 2

Outperformance in exchange-traded fund pricing deviations: Generalised control of data snooping bias

2.1 Introduction

Exchange-Traded Funds (ETFs) are variants of mutual funds that first came to prominence in the early 1990s. ETFs allow market participants to trade index portfolios, similar to how individual investors trade shares of a stock. They seek to track the value and volatility of an underlying benchmark index through the construction of portfolios replicative of the index's constituents. They were first traded on the Toronto Stock Exchange in 1989 and today's market boasts over 1,220 U.S. traded ETFs.[1] Investors seeking ETF outperformance may be tempted to apply a number of performance measures to a large data set of ETFs in order to test for those that are profitable. Given enough tests, they are virtually certain to uncover individually significant ETFs and may näively use these as a basis for portfolio selection decisions. However, in such a set-up, there is a likelihood that these *seemingly significant* outperformers are due to mere chance alone. As the number of simultaneous tests conducted increases so too does the likelihood of such false discoveries. This issue is known as data snooping bias and must be controlled for when studying ETF outperformance. A key contribution in this study is the use of an innovative procedure, proposed in the literature, to control for this problem. The paper further uses an extensive ETF database that offers significant geographic and sector coverage. In this way, the paper provides robust first stage guidance to investors of where inefficiencies may be and, accordingly, where ETFs may provide some investment advantages. The main item of note from the implementation is that, when performance is analysed on a non-risk-adjusted basis only, no funds in the sample are identified as displaying any measure of outperformance.

[1] Investment company institute June 2012 ETF report:
http://www.ici.org/etf_resources/research/etfs_06_12 (Accessed 10/30/12).

It is only the risk-adjusted performance measures that give statistically significant outperformance results and so the insights from these results dominate the commentary. The key takeaways from the study are, firstly, a high proportion of optimised replication, debt asset class, and global/international ETFs exhibit risk-adjusted premiums, highlighting redemption in kind inefficiencies. Secondly, cross-sector and sectoral funds display broadly the same percentage of outperformance. Lastly, high expense ratio and recent inception date ETFs are more likely to exhibit index outperformance, which is of interest to investors seeking to outperform their benchmarks.

The reason for the growth in popularity of ETFs over recent years can be attributed to a number of advantages that they offer over other index-linked products. Tax efficiency and lower expenses are the two most frequently mooted draws for investors, with another being smaller transaction quantities than equivalent futures products, a feature allowing retail investors the opportunity to participate in the market. Empirical studies on active mutual funds have found that, on average, they do not produce above normal returns. Malkiel (1995) and Gruber (1996) show that this inability to beat the market is primarily due to the level of management expenses charged. This performance outcome has increased interest in passive market tracking funds. ETFs aim to replicate index performance but with lower transaction costs and greater tax efficiency than observed in comparable mutual funds. Actively managed ETFs, whose goal is to realise above market returns, only release information on their specific holdings at an end-of-day frequency, whereas the weighted constituents of the passively managed ETFs are always known. Rompotis (2011) cites this as a reason why passive ETFs are advantageous in the eyes of potential arbitrageurs and for their retention as the more popular ETF type. Other miscellaneous strengths of ETFs that have contributed to their rise in popularity have been explicitly identified. Firstly, ETFs provide diversification satisfying broad exposure, be it marketwide or sectoral coverage, with sectoral ETFs facilitating hedging requirements. Secondly, Yu (2005) and Alexander and Barbosa (2008) observe that ETFs do not have short selling restrictions in the same manner as regular stocks so they may be more useful for hedging. Lastly, ETFs are not subject to the uptick rule, which Curcio et al. (2004) suggest as another benefit for shareholders.

A set of 288 U.S. traded ETFs is evaluated in this study using hypothesis tests that seek to identify those that outperform their net asset value (NAV), their underlying index, or a market benchmark. A major contribution to the literature here is the utilisation of a generalised data snooping bias procedure in the ETF performance appraisal setting. Data snooping bias, in this context, is the problem whereby under näive analysis statistically significant outperformance relationships may be identified by pure chance alone. Controlling for data snooping bias is important in order to obtain greater levels of confidence when analysing ETF performance. The false discovery of such random artefacts can greatly mislead an investor's portfolio selection.

Data snooping bias links directly to the broader issue of multiple hypothesis testing in statistical and econometric applications where the issue is commonly referred to as the

multiple comparisons problem. This problem is well addressed in the scientific and medical fields but largely ignored in the empirical finance literature. This paper contributes to the empirical research on ETFs by applying the generalised balanced stepdown procedure of Romano and Wolf (2010), which serves as an improvement over the more conservative seminal reality check bootstrap test of White (2000) and the superior predictive ability test of Hansen (2005). The generalised balanced stepdown procedure of Romano and Wolf (2010) boasts a greater ability to reject false null hypotheses as well as offering balance in the sense that all hypotheses are treated equally in terms of power.

A number of quantitative studies employ such procedures to control for data snooping bias. Sullivan and Timmermann (1999), Hsu and Kuan (2005), Park and Irwin (2007), Marshall et al. (2008) and Qui and Wu (2008) apply the reality check bootstrap test of White (2000) to evaluate the profitability of a wide range of technical trading rules commonly used in industry. Qui and Wu (2008) analyse foreign exchange markets while Marshall et al. (2008) considering a data set of 15 commodities. Hsu et al. (2010) employ a stepwise extension of the superior predictability test of Hansen (2005) to re-evaluate the profitability of technical trading rules, with Bajgrowicz and Scaillet (2012) utilising a false discovery rate (i.e., the proportion of false discoveries to the total number of significant hypothesis tests identified) approach to analyse technical trading rules applied to stock returns. Controlling for data snooping bias in a statistical arbitrage setting, Cummins and Bucca (2012) provide a practical comparison of the stepwise procedure of Romano and Wolf (2007) and the balanced stepdown procedure of Romano and Wolf (2010). They find that the balanced stepdown procedure is unbiased in its approach and is shown to identify many more profitable trading strategies compared to the non-balanced stepdown procedure.

An acknowledgment of this multiple comparisons issue can be seen in both the hedge and mutual fund performance literature but this is not the case for ETFs. In assessing hedge fund performance, Criton and Scaillet (2011) use the false discovery rate to control for data snooping bias. Cuthbertson et al. (2008) and Barras et al. (2010) also utilise the false discovery rate in order to find the proportion of "lucky" mutual funds amongst those with significant individual alphas. However, unless the false discovery rate is zero, it is not possible to identify which of the individual funds are genuinely outperforming. This study significantly extends this literature, incorporating the more recent balanced stepdown procedure of Romano and Wolf (2010) and applying this in the ETF realm to identify both individual ETFs and ETF cohorts that outperform. The Romano and Wolf (2010) procedure further works on the generalised familywise error rate rather than the false discovery rate — the former being the actual number of false discoveries from the set of all true hypotheses.

The methodologies used in previous ETF studies raise concerns around the validity of the inferences drawn, insofar as they lack data snooping bias controls and in many cases conduct less formal hypothesis tests. This can greatly mislead an investor's portfolio selection. Addressing this issue is important as it calls into question, and potentially

undermines, the findings and conclusions in the literature. The major argument of this paper is therefore that in order for one to be robustly confident of one's ETF performance conclusions, one must control for the multiple comparisons problem. The robustness of one's economic arguments is intrinsically linked to the robustness of the econometric analysis. This requires a fundamental shift in the way that ETF performance is analysed econometrically; a fundamental shift that is equally required in the mainstream empirical finance literature. The majority of research conducted to date has centred on data sets comprising small numbers of large ETFs, single ETF families or industries, with measurements being applied inconsistently across the differing studies, inhibiting effective cross comparison. This body of work amends that, primarily through the use of a large, diverse sample size, which incorporates many sectoral and internationally focused indices. We investigate numerous ETF attributes and their ability to dictate outperformance, alongside including a recent time period. The effect of replication type and asset class focus on ETF performance for instance has not been rigorously tested in the literature and as such this study incrementally contributes in this way. This work may be of interest to a variety of stakeholders. Firstly, investigating ETF outperformance is significant from an academic perspective as it furthers our understanding of the market's pricing dynamics. Secondly, the wider investment community would benefit from the work as an aid in identifying specific ETF cohorts suitable to individual portfolio requirements. Lastly, despite it not being the focus of this study, the data snooping bias issues raised offer broader insights to arbitrageurs by emphasising the importance of controlling for data snooping bias in order to robustly identify mispricings and trading signals.

The remainder of the paper is organised as follows. In Section 2.2 we discuss in-kind deviations along with performance differences between ETF prices, underlying indices, and a market benchmark. In section 5.3.4 we discuss the issue of data snooping bias and link this to the broader issue of multiple hypothesis testing. We also discuss the details of the balanced stepdown procedure of Romano and Wolf (2010), along with the associated operative method that allows for computational efficiency. The empirical analysis is outlined in Section 2.4, where we describe the data set and define the formal hypothesis tests. Section 2.5 presents the results of the empirical analysis and considers various attributes of outperforming funds. Section 5.5 concludes.

Summary of contributions

- What ETFs display a Net Asset Value premium?
- What ETFs outperform their underlying index?
- What ETFs outperform market wide benchmarks?
- How do specific groupings of ETFs differ in terms of outperformance?

2.2 Outperformance

We examine ETF outperformance on three levels: ETF NAV premium; ETF price versus its tracked underlying index; and ETF price versus a market return benchmark. NAV premium refers to the amount that the secondary market price of the ETF trades above its calculated NAV. If the amount is negative, it is referred to as a NAV discount. The creation/redemption/deletion procedure facilitates exploitation in such situations, whereby the investor can exchange units of trust for the underlying index's stock and vice versa. The return to optimal Law of One Price levels would occur if there were no limits to arbitrage, with the most notable observed limitations being market frictions (redemption fees and bid-ask spreads). There is empirical evidence of an inconsistency in premium levels between domestic and non-domestic funds, whereby non-domestic funds display persistent premiums with U.S. domestic funds tracking their NAVs relatively well. Elton et al. (2002) and Ackert and Tian (2008) both observe that U.S. ETFs are priced close to NAVs, while Jares and Lavin (2004) and Engle and Sarkar (2006) report that some country ETFs display premiums/discounts. Elton et al. (2002) report an average annual return from holding Spiders[2] of 21.91% between the years 1994 and 1998, with the NAV return being slightly lower at 21.89%. However they highlight, that the figures may overstate the true difference as Spiders continue to trade for up to 15 minutes after the New York Stock Exchange closes. Engle and Sarkar (2006) use both daily and intra-day data to investigate short-term deviations between the traded price and NAVs of 21 domestic (U.S.) and 16 international ETFs between April and September 2000. They find that ETFs trade in a premium range of between -0.1 bps and 4.6 bps. U.S. ETFs show minute premiums that are smaller than typical bid-ask spreads whereas international ETFs are less accurately priced due to higher tax and creation/redemption costs. Jares and Lavin (2004) consider mispricings in two Asian ETFs, namely Hong Kong and Japan country funds. They conclude that the non-synchronised trading hours between the U.S. and foreign markets induces the presence of premiums. This study incorporates ETFs from both of these geographic locations.

An ETF is said to have an index tracking error if a fund does not perfectly mirror its underlying benchmark index. Elton et al. (2002) find that Spiders underperform the S&P 500 Index by 28 bps, the two main causes for this underperformance being the management fee of 18 bps and the dividend being placed in a non-interest bearing account, which results in another 9.95 bps loss. The influence of expense ratio on ETF outperformance is one of the many factors addressed in Section 2.5. Harper et al. (2006) provide a comparison of ETFs and closed-end country funds (CEFs), observing no significant tracking error between iShares ETFs and MSCI[3] indices from April 1996 to December 2001. DeFusco et al. (2011) study the three most liquid ETFs, the Spiders, Diamonds, and Cubes.[4] Through setting

[2]Standard & Poor's Depository Receipts ("Spiders" or SPDRs) track the performance of the S&P 500 Index.

[3]MSCI is an abbreviation of Morgan Stanley Capital International. iShares are ETFs tracking the performance of MSCI individual country market indices.

[4]Diamonds and Cubes are ETFs tracking the performance the Dow Jones Industrial Average and NAS-

out five hypothesis tests on the non-synchronous price deviations between the ETFs and the notional price of the index, they conclude that the tracking error is a non-zero, non-normal, stationary process that is dependent on both the accumulation of dividends and on the size of the benchmark index. This paper deals with the size issue through the proxy of each ETF's total assets under management.

Market tracking error in this context refers to how much an ETF under/outperforms a broad market index. The majority of mutual fund and ETF studies to date utilise the S&P 500 as their U.S. benchmark index proxy alongside incorporating risk-adjusted returns into the analysis. Phengpis and Swanson (2009), using monthly data and incorporating the Wilshire 3000 Index to represent the U.S. market return, find that country iShares are not heavily exposed to U.S. market risk. The results are obtained using a new two factor test specification with the iShares typically mirroring their underlying market indices up to the end of March 2007. The relationship between a U.S. market benchmark and country iShares is revisited in this study. Mateus and Kuo (2008) also study ETF performance, providing a comparative analysis of 20 country-specific ETFs with the S&P 500 Index over a five-year period. Risk-adjusted measures are used, namely, Sharpe, Treynor, and Sortino ratios. Sharpe and Sortino ratios are again calculated by Rompotis (2011), who shows that the majority of the 50 selected iShares in his sample outperform the S&P 500 on both an annual and aggregate basis over the 2002 to 2007 period.

It is necessary to briefly highlight some methodological deficiencies contained in the above papers. These deficiencies call into question some of the economic reasoning put forward. A common adjustment method is used to assess the scale of the multiple comparisons problem in the literature. This adjustment is the Bonferroni correction which involves controlling the familywise error rate (i.e., the probability of obtaining one or more false discoveries) by using for each hypothesis test a per comparison cut-off value ($\tilde{\alpha}$) equal to the required significance level (α) divided by the number of hypothesis tests n, i.e., $\tilde{\alpha}=\alpha/n$.

Firstly, Ackert and Tian (2008) fail to conduct formal hypothesis tests on the observed premium, simply reporting 5% and 95% intervals for the 28 individual ETFs. Such a naïve approach fails to control for the multiple comparisons problem. Even without accounting for the multiple comparisons issue, none of the observed premiums are statistically significant, yet inferences from the tests are used as the basis for the paper's primary contribution—to uncover a U-shape between illiquidity and fund premium. Harper et al. (2006) promote the use of international ETFs over CEFs in an analysis of passive versus active investment strategies. In doing so, a test is conducted whereby Sharpe ratios are calculated for both the passive ETFs and the active CEFs with t-tests of the Jobson Korkie statistic reported; 29 simultaneous hypotheses are tested about a data set with rejections listed at 5% and 10% significance levels. Four passive ETFs are said to significantly outperform the active CEFs. Using a simplified correction methodology for illustrative purposes, a Bonferroni adjustment would lead to lowering an alpha level of 10% to 0.34% (10%/29

DAQ 100 indices, respectively.

= 0.34%). None of these four supposedly significant ETFs genuinely outperform the CEFs after this adjustment is made. Jares and Lavin (2004) find that premiums in Japan and Hong Kong ETFs are positively related to subsequent ETF returns and propose a trading strategy to exploit this. Comprehensive economic reasoning for this observed dynamic is not given. One-tailed *t*-test statistics are given for the hypothesis that the proposed trading rule exceeds a buy and hold strategy, indicating significance at the 2.5% and 1% levels for Japan and Hong Kong, respectively. There are 12 (6 years by 2 ETFs) simultaneous hypotheses tested in the paper, resulting in a Bonferroni adjustment that decreases a 1% significance level to 0.08% (1%/13 = 0.08%). There is insufficient information to comment on the genuine significance of the Hong Kong ETF (it is simply listed as being significant at the 1% level) but under such an adjustment the profitable trading strategy associated with the Japan ETF would be classed as a false discovery. Jares and Lavin (2004) test only two ETFs, yet generalisations to all foreign ETFs are tentatively made. Another major issue with the analysis is that it is devised and tested in the same sample period of 1996-2001. Jares and Lavin (2004) claim that the results uncovered are "almost too good to be true" and with such flawed in-sample evaluation, this is quite likely to be the case. A much more rigorous and robust trading strategy analysis, which controls for the data snooping bias, but within an energy market setting, is that of Cummins and Bucca (2012).

Elton et al. (2002) examines the performance of the Standard and Poor's Depository Receipts (SPDRs). No formal hypothesis tests are conducted to evaluate the significance of the tracking errors reported. Instead a difference is simply taken between the levels of the index and ETF and a premium level frequency distribution constructed to form its inferences. DeFusco et al. (2011) propose five hypotheses; namely that ETF tracking error is (i) a stationary process, (ii) normally distributed, (iii) zero, (iv) linked to dividend accumulation, and (v) that indices with fewer stocks display smaller tracking errors. Utilising a sample of just three U.S. ETFs, the tracking errors are tested on days when dividends are paid; when zero, it is inferred that dividends accumulated affect the size of the tracking error. This hypothesis is rejected for the log price of the SPY index at the 5% level and for the prices SPY index at the 10% level. However no control for the multiple comparisons issue is in place here. A family of six simultaneous hypotheses result in a Bonferroni correction of 10%/6 = 1.667% and 5%/6 = 0.833%, leading to adjusted alphas of 1.67% and 0.83% for 10% and 5% significance levels, respectively. Under this adjustment, there would be no hypothesis rejections.

2.3 Multiple hypothesis testing: data snooping bias

The objective of the study is to formally and statistically test for the presence of outperformance in ETF markets. This will inevitably involve the testing of a large number of performance measure implementations simultaneously. In particular, 11 pricing deviations are considered for each of the 288 ETFs, leading to the simultaneous assessment of 3,168 performance measures. This introduces the well-established issue of data snooping bias,

which in this context, is the problem whereby under näive analysis, statistically significant outperformance relationships may be identified by pure chance alone. The false discovery of such random artefacts can greatly mislead an investor's portfolio selection and links directly to the broader issue of multiple hypothesis testing in statistical and econometric applications.

The issue with multiple hypothesis testing is that the probability of false discoveries, i.e., the rejection of true null hypotheses by chance alone, is often significant. There are a number of approaches described in the literature to deal with this multiple comparisons problem and control for the familywise error rate (FWER) and related variants. Romano et al. (2010) provide an excellent summary of the issues and the literature. The FWER is defined as the probability that at least one or more false discoveries occur. Consistent with the notation of Romano et al. (2010), the following definition is made:

$$FWER_\theta = P_\theta \left\{\text{reject at least one null hypothesis } H_{0,s} : s \in \mathcal{I}(\theta)\right\},$$

where $H_{0,s}, s = 1, \ldots, S$, is a set of null hypotheses; and $\mathcal{I}(\theta)$ is the set of true null hypotheses. Controlling the FWER involves setting a significance level α and requiring that $FWER_\theta \leq \alpha$. This approach is particularly conservative given that it does not allow even for one false discovery and so is criticised for lacking power, where power is loosely defined as the ability to reject false null hypotheses, i.e., identify true discoveries (Romano et al. 2010). The greater S, the more difficult it is to make true discoveries.

To deal with this weakness, generalised FWER approaches have been proposed in the literature. The generalised FWER seeks to control for k (where $k \geq 1$) or more false discoveries and, in so doing, allows for greater power in multiple hypothesis testing. The generalised k-FWER is defined as follows:

$$k\text{-}FWER_\theta = P_\theta \left\{\text{reject at least } k \text{ null hypothesis } H_{0,s} : s \in \mathcal{I}(\theta)\right\}.$$

Towards building a framework to identify outperforming ETFs, with statistical significance, the following one-sided hypothesis test is considered:

$$H_{0,s} : \theta_s \leq 0 \quad \text{vs.} \quad H_{1,s} : \theta_s > 0.$$

The objective is to control for the multiple comparisons in this scenario through the generalised FWER, which offers greater power while also implicitly accounting for the dependence structure that exists between the tests. Before outlining the balanced stepdown procedure of Romano and Wolf (2010), it is first necessary to present the (inferior) single-step procedure designed around the generalised FWER. The advantages of the Romano and Wolf (2010), procedure are better appreciated with this context.

2.3.1 Single-step procedure

Assume a set of test statistics $T_{n,s} = \hat{\theta}_{n,s}$ associated with the hypothesis tests, where n

denotes the sample size of the data used for estimation. Letting $A \equiv \{1, \ldots, S\}$, the single-step procedure proceeds by rejecting all hypotheses where $T_{n,s} \geq c_{n,A}(1-\alpha, k)$, and where $c_{n,A}(1-\alpha, k)$ represents the $(1-\alpha)$-quantile of the distribution of k-max $\left(\hat{\theta}_{n,s} - \theta_s\right)$ under P_θ. With P_θ unknown, the critical value $c_{n,A}(1-\alpha, k)$ is also unknown. However, an estimate critical value may be determined using appropriate bootstrapping techniques. That is, the critical value $\hat{c}_{n,A}(1-\alpha, k)$ is estimated as the $(1-\alpha)$-quantile of the distribution of k-max $\left(\hat{\theta}^*_{n,s} - \hat{\theta}_{n,s}\right)$ for $\hat{P}_\theta$ an unrestricted estimate of P_θ. See Romano et al. (2010) for further technical details.

2.3.2 Balanced stepdown procedure

The single-step procedure is improved upon with the balanced stepdown procedure of Romano and Wolf (2010) by allowing for subsequent iterative steps to identify additional hypothesis rejections. It also offers balance by construction in the sense that each hypothesis is treated equally in terms of power. The stepdown procedure is constructed such that at each stage, information on the rejected hypotheses to date is used in re-testing for significance on the remaining hypotheses.

Again assume a set of test statistics $T_{n,s} = \hat{\theta}_{n,s}$ associated with the hypothesis tests, where n is again the sample size of the data used for estimation. Introducing some notation, let $H_{n,s}(\cdot, P_\theta)$ denote the distribution function of $\left(\hat{\theta}_{n,s} - \theta_s\right)$ and let $c_{n,s}(\gamma)$ denote the γ-quantile of this distribution. The confidence interval

$$\left\{\theta_s : \hat{\theta}_{n,s} - \theta_s \leq c_{n,s}(\gamma)\right\}$$

then has coverage probability γ. Balance is the property that the marginal confidence intervals for a population of S simultaneous hypothesis tests have the same probability coverage. Within the context of controlling the generalised k-FWER, the overall objective is to ensure that the simultaneous confidence interval covers all parameters $\theta_s, s = 1, ..., S$, except for at most $(k-1)$ of them, for a given limiting probability $(1-\alpha)$, while at the same time ensuring balance (at least asymptotically). So, what is sought is that

$$P_\theta\left\{\hat{\theta}_{n,s} - \theta_s \leq c_{n,s}(\gamma) \quad \text{for all but at most } (k-1) \text{of the hypotheses}\right\}$$

$$\equiv P_\theta\left\{H_{n,s}\left(\hat{\theta}_{n,s} - \theta_s, P_\theta\right) \leq \gamma \quad \text{for all but at most } (k-1) \text{of the hypotheses}\right\}$$

$$\equiv P_\theta\left\{k\text{- max}\left(H_{n,s}\left(\hat{\theta}_{n,s} - \theta_s, P_\theta\right)\right) \leq \gamma\right\} = 1 - \alpha.$$

Letting $L_{n,\{1,\ldots,S\}}(k, P_\theta)$ denote the distribution of k- max $\left(H_{n,s}\left(\hat{\theta}_{n,s} - \theta_s, P_\theta\right)\right)$, the appropriate choice of the coverage probability γ is then $L^{-1}_{n,\{1,\ldots,S\}}(1-\alpha, k, P_\theta)$.

Given that P_θ is unknown, it necessary to use appropriate bootstrapping techniques

to generate an estimate of the coverage probability $L^{-1}_{n,\{1,\ldots,S\}}\left(1-\alpha,k,\hat{P}_\theta\right)$, under $\hat{P}_\theta$. Therefore, from this development it is possible to define the simultaneous confidence interval

$$\left\{\theta_s : \hat{\theta}_{n,s} - \theta_s \leq H^{-1}_{n,s}\left(L^{-1}_{n,\{1,\ldots,S\}}\left(1-\alpha,k,\hat{P}_\theta\right),\hat{P}_\theta\right)\right\}.$$

The right-hand side of the above inequality will form the basis of the critical value definitions used within the stepdown procedure. See Romano and Wolf (2010) for further technical details. Note that although the above development was made assuming the full set of hypothesis tests, it equally applies to any subset $K \subseteq \{1,\ldots,S\}$. Hence, the balanced stepwise algorithm may now be described as follows.

- **Step 1:** Let A_1 denote the full set of hypothesis indices, i.e. $A_1 \equiv \{1,\ldots,S\}$. If for each hypothesis test, the associated test statistic $T_{n,s}$ is less than or equal to the corresponding critical value estimate, $\hat{c}_{n,A_1,s}(1-\alpha,k) \equiv H^{-1}_{n,s}\left(L^{-1}_{n,A_1}\left(1-\alpha,k,\hat{P}_\theta\right),\hat{P}_\theta\right)$, then fail to reject all null hypotheses and stop the algorithm. Otherwise, proceed to reject all null hypotheses $H_{0,s}$ for which the associated test statistics exceeds the critical value level, i.e., where $T_{n,s} > \hat{c}_{n,A_1,s}(1-\alpha,k)$.

- **Step 2:** Let R_2 denote the set of indices for the hypotheses rejected in Step 1 and let A_2 denote the indices for those hypotheses not rejected. If the number of elements in R_2 is less than k, i.e., $|R_2| < k$, then stop the algorithm, as the probability of k or more false discoveries is zero in this case. Otherwise, the appropriate critical value to be applied for each hypothesis test s at this stage is calculated as follows:

$$\hat{d}_{n,A_2,s}(1-\alpha,k) = \max_{I\subseteq R_2,|I|=k-1}\left\{\hat{c}_{n,K,s}(1-\alpha,k) : K \equiv A_2 \cup I\right\}.$$

  Hence, additional hypotheses from A_2 are rejected if $T_{n,s} > \hat{d}_{n,A_2,s}(1-\alpha,k), s \in A_2$. If no further rejections are made then stop the algorithm.
  $\vdots$

- **Step j:** Let R_j denote the set of indices for the hypotheses rejected up to Step $(j-1)$ and let A_j denote the indices for those hypotheses not rejected. The appropriate critical value to be applied for each hypothesis test s at this stage is calculated as follows:

$$\hat{d}_{n,A_j,s}(1-\alpha,k) = \max_{I\subseteq R_j,|I|=k-1}\left\{\hat{c}_{n,K,s}(1-\alpha,k) : K \equiv A_j \cup I\right\}.$$

  Hence, additional hypotheses from A_j are rejected if $T_{n,s} > \hat{d}_{n,A_j,s}(1-\alpha,k), s \in A_j$. If no further rejections are made then stop the algorithm.
  $\vdots$

At each step j in the stepwise procedure, the hypotheses that are not rejected thus far are re-tested over a smaller population of hypothesis tests than previously. The size of this smaller population is given $(|A_j| + k - 1)$, which includes all the hypotheses within A_j, in addition to $(k-1)$ hypotheses drawn from those hypotheses already rejected, i.e., drawn from R_j. Given that control of the generalised k-FWER is the premise of the procedure, it is expected that there are at most $(k-1)$ false discoveries amongst the set of hypotheses rejected R_j. However, it is not known which of the rejected hypotheses may represent false discoveries. Hence, it is necessary to circulate through all combinations of R_j, of size $(k-1)$, in order to obtain the appropriate critical values. A maximum critical value $\hat{d}_{n,A_j,s}(1-\alpha,k)$ must be determined for each hypothesis test s. This adds an additional layer of computational burden on the algorithm.

2.3.2.1 Operative method

In requiring to circulate through all subsets of R_j, of size $(k-1)$, in order to obtain the maximum critical value to apply at each stage of the stepdown procedure, the algorithm can become highly, if not excessively, computationally burdensome. Depending on the $|R_j|$ and the value of k, the number of combinations ${}^{|R_j|}C_{k-1}$ can become very large. Romano and Wolf (2010) therefore suggest an operative method that reduces this computational burden, while at the same time maintaining much of the attractive properties of the algorithm.[5]

It is first necessary to be able to order the hypothesis tests rejected up to step $(j-1)$ in terms of significance. To this end, it is noted that marginal p-values can be obtained as follows:

$$\hat{p}_{n,s} \equiv 1 - H_{n,s}\left(\hat{\theta}_{n,s}, \hat{P}_\theta\right).$$

This gives the following ascending order for the significance of the hypothesis tests:

$$\hat{p}_{n,r_1} \leq \hat{p}_{n,r_2} \leq \ldots \leq \hat{p}_{n,r_{|R_j|}},$$

where $\left\{r_1, r_2, \ldots, r_{|R_j|}\right\}$ is the appropriate permutation of associated hypothesis test indices that gives this ordering. As before, a maximum number of combinations, N_{max}, at each step of the algorithm is defined. Then an integer value M is chosen such that ${}^{M}C_{k-1} \leq N_{max}$, leading to the calculation of the critical values as follows:

$$\hat{d}_{n,A_j,s}(1-\alpha,k) = \max_{I \subseteq \left\{r_{\max\left(1,\left|R_j\right|-M+1\right)},\ldots,r_{\left|R_j\right|}\right\}, |I|=k-1} \left\{\hat{c}_{n,K,s}(1-\alpha,k) : K \equiv A_j \cup I\right\}.$$

What this serves to do is to replace circulating through all the hypothesis tests rejected to date with that of circulating through only the M least significant hypothesis tests

[5] Attractive properties include conservativeness, which allows for finite sample control of the k-FWER under P_θ, and provides asymptotic control in the case of contiguous alternatives Romano and Wolf (2007).

rejected. Of course, in the case where $M \geq |R_j|$, then this amounts to circulating through all the hypotheses rejected. Although this approach is premised on the assumption that the (up to $k-1$) false discoveries lie within the least significant hypotheses rejected so far, it does offer significant computational efficiencies for the algorithm. It is this operative method that is used for the empirical analysis in subsequent sections of this chapter, as well as in Chapters 4 and 5.[6]

2.4 Empirical analysis: framework and data

The balanced stepdown procedure described in the previous section offers a more generalised and flexible approach to controlling data snooping bias than previous methodologies in the literature. In particular, it controls the generalised FWER using a superior stepwise procedure that offers balance by construction. This property of balance ensures that each outperformance measure is treated equally in terms of power, i.e., the ability to reject false null hypotheses, and so outperformance measures with large deviations do not dominate those with lower deviations. This is one of the key motivations for using the balanced stepdown procedure for the empirical analysis of this study. Firstly, in order to test for ETF premiums, the differences between the mean daily log return of the quoted ETF price and the mean daily log return of its reported NAV are examined, with the null hypothesis being that the ETF return is less than or equal to the NAV return, i.e., no outperformance.[7] The analysis is extended through the implementation of traditional risk-adjusted measures such as the Sharpe, Sortino, and Treynor ratio test statistics with the null hypotheses of no outperformance again in place. The same approach is employed in constructing index and market outperformance hypothesis tests, replacing the NAV series with the fund's underlying index and the S&P 500 series respectively.

The three risk-adjusted ratios are now examined. The Sharpe ratio (Sharpe 1966), is the most commonly used ex post measure of risk-adjusted performance in the ETF literature. It is a measure of an investment's performance per unit of risk, whereby standard deviation is used as a proxy for the portfolio's risk. The Treynor ratio is a variant of the Sharpe ratio that incorporates a CAPM-based excess return component, effectively giving excess return per unit of market risk. Where the normality assumption is not in place for returns, it is beneficial to consider the Sortino ratio, the third risk-adjusted measure of performance considered. It is again based on the Sharpe Ratio but differentiates between upside and downside risk whereby it does not penalised for upside volatility. Formally, these risk-adjusted measures are summarised as follows:

$$\rho_p = \frac{R_p - r_f}{\eta_p}$$

,

[6] The resampling based and p-value based MHT algorithms were made available to me by Dr Mark Cummins.

[7] Use of the log return methodology is in line with Engle and Sarkar (2006).

where ρ_p = portfolio's Sharpe, Sortino or Treynor ratios, R_p = portfolio return, η_p = standard deviation of portfolio for Sharpe, standard deviation of negative returns for Sortino or market beta for Treynor ratios and r_f = risk-free rate.[8]

As referred to previously, for each of the 288 ETFs, 11 pricing deviations are calculated on a daily basis.[9] To complete the set-up of the empirical analysis, it is necessary to discuss the choice of generalising parameter k and the probability parameter α to be used within the balanced stepdown procedures. To ensure tight control of the number of false discoveries while at the same time offering power to the tests, k is chosen to ensure that no more than 1% of the tests represent false discoveries. The significance level α chosen is 5% alongside an N_{max} value of 100 combinations in line with Romano and Wolf (2010).

The data set comprises 288 U.S. domiciled equity, commodity, and debt ETFs with pre-2008 inception dates. The period of study is 2008-2012, a time span that is chosen to strike an acceptable balance between being sufficiently long to retain power in the proposed econometric tests and recent enough to be representative of the vast array of ETFs. Data on end-of-day market price, reported NAV, and the notional value of the tracked is downloaded from Bloomberg for each fund. Supplementary data on total asset value, underlying asset class, replication strategy, expense ratio, industry and country focus is also gathered. Table 2.1 provides the cohort proportions of the data set. It includes funds in the assets under management range of \$9.72 million to \$101,187.40 million with a broad industry split; 18 from the energy sector, 14 from technology, 12 from financial services, and 11 from health and biotechnology, for example.[10] The median expense ratio is 0.51, with a range of 0.09 to 2.55. The sample includes both many U.S. and non U.S. focused funds,[11] along with full, optimised, and derivative replication types. A major contribution of this study is borne out of the inclusion of these additional factors as they allow for more informed portfolio selection decisions. Average daily risk free rates are downloaded from the website of Kenneth French[12] in a manner similar to Rompotis (2011). These are to be utilised in the calculation of risk-adjusted performance measures.

As identified earlier, the use of the Sortino ratio is appropriate and valid where returns are shown to be non-normal. For completeness, the normality of returns is formally tested for each of the 288 ETF price, the 288 NAV and the 288 index series. The hypothesis that the returns are normal is tested using the Jarque-Bera two-sided goodness-of-fit test.[13]

[8] A wealth of alternative risk measures exist in the portfolio management literature, some of which may lead to different results and distinct inferences being drawn. However, we chose to adopt those used by both Mateus and Kuo (2008) and Rompotis (2011).

[9] Note that the construction of the Treynor ratio, which incorporates the market beta, is the reason for the omission of a Mkt TE TR measure. TE is "tracking error".

[10] "No Industry Focus Given" is used to denote sector ETFs where no industry focus has been provided by Bloomberg.

[11] International ETFs refer to investments targeted at multiple geographic locations outside of the home market (U.S.) whereas global ETFs refer to investments targeted at multiple geographic locations inclusive of the home market (U.S.).

[12] Kenneth French's website:
http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (accessed 06/30/12).

[13] The null hypothesis is that the deviations are normally distributed with unspecified mean and standard deviation, whereas the alternative is that the deviations are not normally distributed.

The multiple comparisons problem presents itself here again due to conducting 864 Jarque-Bera normality tests simultaneously. Given the availability of p-values from the Jarque-Bera tests, the use of a p-value-based multiple hypothesis testing (MHT) procedure is appropriate here.[14] The MHT framework of Romano and Shaikh (2006) used that controls for what is referred to as the false discovery proportion (FDP). It is defined as:

$$FDP \equiv \begin{cases} \frac{FR}{TR}, & TR > 0 \\ 0, & TR = 0 \end{cases},$$

where FR denotes the number of false rejections and TR denotes the total number of rejections. Romano and Shaikh (2006) propose a stepdown procedure that controls the FDP, whereby for a given proportion $\tilde{\gamma}$ and significance level $\tilde{\alpha}$,

$$P\left\{FDP > \tilde{\gamma}\right\} \leq \tilde{\alpha}$$

.

For the set of hypothesis tests $H_{0,i}, i = 1, \ldots, 864$, there are available p-values, $\hat{p}_i, i = 1, \ldots, s$. The p-values are ordered from the most significant to the least significant, i.e., $\hat{p}_{(1)} \leq \hat{p}_{(2)} \ldots \leq \hat{p}_{(s)}$, and the associated ordered null hypotheses $H_{0,(i)}$ are rejected if and only if $\hat{p}_{(i)} \leq \tilde{\alpha}'_{(i)}$ with the cut-off values defined as:[15]

$$\tilde{\alpha}'_{(i)} \equiv \tilde{\alpha}_{(i)}/C,$$

where

$$\tilde{\alpha}_{(i)} = \frac{(\lfloor \tilde{\gamma} i \rfloor + 1)\tilde{\alpha}}{s + \lfloor \tilde{\gamma} i \rfloor + 1 - i}$$

and

$$C \equiv C(\tilde{\gamma}, \tilde{\alpha}, s) = \max_{|I|} S(\tilde{\gamma}, \tilde{\alpha}, |I|),$$

$$S(\tilde{\gamma}, \tilde{\alpha}, |I|) \equiv |I| \sum_{j=1}^{N} \frac{\beta_j - \beta_{j-1}}{j}$$

,

$$N \equiv N(\tilde{\gamma}, \tilde{\alpha}, |I|) = \min\left\{ \lfloor \tilde{\gamma} s \rfloor + 1, |I|, \left\lfloor \tilde{\gamma} \left(\frac{s - |I|}{1 - \tilde{\gamma}} + 1 \right) \right\rfloor + 1 \right\},$$

and where

[14] There are two classifications of procedure identified in the MHT literature: (i) re-sampling-based and (ii) p-value-based. The balanced stepdown procedure outlined in Section 4 is of the re-sampling type, involving a bootstrapping component. See Romano and Wolf (2010) for more details on both classifications.

[15] It is important to emphasise the subtle difference in notation. $H_{0,i}$ is the i-th hypothesis test considered and $\hat{p}_i$ is the associated p-value. In contrast, $H_{0,(i)}$ is used to denote the i-th hypothesis when all hypotheses are ordered in terms of significance from the most significant to the least significant, with $\hat{p}_{(i)}$ denoting the associated ordered p-value.

$$\beta_0 \equiv 0,$$

$$\beta_m \equiv \frac{m}{max\left\{s+m-\left\lceil\frac{m}{\tilde{\gamma}}\right\rceil+1,|I|\right\}}, m=1,...,\lfloor\tilde{\gamma}s\rfloor,$$

and

$$\beta_{\lfloor\tilde{\gamma}s\rfloor+1} \equiv \frac{\lfloor\tilde{\gamma}s\rfloor+1}{|I|}.$$

This approach boasts robustness to the dependence structure of the p-values. The proportion parameter $\tilde{\gamma}$ is chosen to be 5% with the significance level $\tilde{\alpha}$ set at 5% also. See Romano and Shaikh (2006) for further details.

Upon implementing the procedure, significant non-normality is observed for all price, NAV, and underlying index series, confirming the use of the Sortino ratio as appropriate. Even though the sample ETF returns are not normally distributed, traditional risk-adjusted ratios, Sharpe and CAPM-based Treynor ratios are extensively used in previous studies and this study as well. They provide an intuitive way of comparing results between studies and offer numerous practical applications in measuring both ETF and mutual fund performance (Mateus and Kuo 2008). Plantinga et al. (2001) examine the application of risk-adjusted ratios to Euronext mutual funds and find that there is a high correlation between the classic Sharpe ratio and a ratio controlling for downside risk, adding further weight to the applicability of such performance measures. The next section presents the results subsequent to applying the balanced stepdown procedure described in Section 5.3.4 to the data set.

2.5 Empirical analysis: results

The results of the operative balanced stepdown procedure of Romano and Wolf (2010) are presented in Figure 2.1, giving the percentage (the actual numbers are given in parentheses) of ETFs in the sample that show specific outperformance measures. The main item of note is that none of the log return outperformance measures are significant under the balanced stepdown procedure. This leads to relying primarily on inferences made around the risk-adjusted measures for the remainder of the paper. The various measures display differing numbers of outperforming funds; for instance, 56 funds show market benchmark outperformance under the Sharpe ratio with almost twice that figure, 105 funds, outperforming the market under the Sortino ratio measure. Summary statistics for the significant outperforming funds are given in Table 2.3, providing the average outperformance measure. The results highlight the importance of controlling for data snooping bias. On the basis of the three non-risk-adjusted measures, i.e., premium, index tracking error and market tracking error, none of the funds outperform. Failure to apply the data snooping bias control procedure would have led to the näive identification of outperformance and so investing

on such a basis would constitute näive and misinformed fund.

A number of ETF attributes are now analysed to determine what classes of ETFs are most likely to demonstrate risk-adjusted outperformance and specifically what outperformance measures they show. Geographic and industry focus are the first to be considered. The geographic focus of ETFs is studied in Figure 2.2, with a high proportion of global, international and other focused funds showing some measure of outperformance. U.S. focused funds on the other hand show a lower proportion of outperformance, although in absolute terms of course the number of funds outperforming is higher at 118. Risk-adjusted premium is a primary driver of these results, as seen in Table 2.4; 63% and 79% of global and international ETFs respectively, show premium Sharpe ratio outperformance, with only 10% for U.S. funds. These findings are in line with Jares and Lavin (2004) and Engle and Sarkar (2006) who also observe premiums among a high percentage of foreign ETFs, and Elton et al. (2002) and Ackert and Tian (2008), who document a low proportion of U.S. focused funds displaying premiums. A lack of synchronisation between NAV calculations and underlying market closes coupled with the time-snap used for exchange rate conversions are often cited reasons for the presence of premiums in ETFs focused over multiple countries/time zones. Further to this, country-specific trading taxes, prohibitions on transactions made by foreigners, longer delivery periods, and greater price risk in assembling packages, all reduce the ability to hedge such positions and increase the likelihood of premiums. In contrast to this, liquidity, latency advantages, and reduced market frictions may allow for easier exploitation of deviations among U.S. focused ETFs.

Figure 2.3 graphs the percentage of ETFs exhibiting some measure of outperformance, split by industry focus. The main item of interest is the comparison of cross-sectoral and single-sector funds. The proportion of funds in each group displaying outperformance is almost identical at 74% and 72%,[16] for cross-sectoral and single-sector funds, respectively. This indicates that inefficiencies are as likely/unlikely to appear in either category. A more in-depth breakdown of the specific industries is also available.

Relatively high percentages of energy, precious metals and real estate ETFs exhibit outperformance with lower numbers observed for financial services ETFs. The high proportion of outperformance observed for these funds are borne out of market TE Sharpe ratios, as shown in Table 2.5, indicating that 56%, 71%, and 50% of energy, precious metals and real estate ETFs, respectively, outperform the market. Precious metals became a safe haven for investors due to the poor performance of equities over the turbulent 2008-2012 period, with the energy sector being buoyed by increased manufacturing demand from China and real estate benefiting from a global decrease in the cost of capital. Financial services in contrast register no ETFs outperforming the market, primarily due to the credit crisis of 2007-2009 and its regulatory legacy.

The next attributes analysed are the assets each ETF attempts to replicate and how they conduct the replication. Full replication is the most widely employed strategy in the data set but only 68% of its funds exhibit outperformance, as shown in Figure 2.4.

[16]Seventy-two percent is an aggregation of the individual sector funds results.

In comparison, 29 ETFs pursuing derivative replication display at least one significant outperformance measure, equating to 83% of its sample. They do not incur the same level of transaction fees as other strategies, which induces outperformance; however, they do house greater counterparty credit risk concentration, which is significantly re-evaluated over the period. Table 2.6 gives an insight into what outperformance measures are seen in these groups. The main item of note is the presence of significant premium outperformance and an absence of significant market outperformance among optimised ETFs, with 50% of optimised funds having a significant Sharpe ratio premium, in contrast to just 11% showing Sharpe ratio market outperformance. An optimised replication strategy involves constructing a portfolio that is a representative subset of the underlying index when full replication of an index's constituents is not possible, be it for cost, liquidity, or regulatory reasons. Such trading impediments, alongside a reliance on historical data to select an index's representative constituents can lead to redemption in kind inefficiencies.

In relation to asset class, the majority of ETFs in the data set have an equity focus; 263 out of 288 (91%). The prevalence of outperformance is broadly in line with these proportions, as seen in Figure 2.4. When looking at the small number of non-equity ETFs in the sample, it can be seen that all of the asset allocation-focused and almost 90% of the debt-focused funds register significant outperformance measures. A significant Sharpe ratio premium is observed for 78% of debt funds with asset allocation-focused ETFs demonstrating index outperformance, according to Table 2.7. Such inefficiencies reside in debt ETFs due to a lack automated trading in the asset class as transactions occur primarily in large blocks between trading desks and institutional clients. Asset allocation-focused ETFs comprise a diverse range of underlyings, which makes exact replication costly.

The final attributes to be examined are the size of the ETF, how much it costs, and when it is first traded. Table 2.8 demonstrates what particular cohorts are most likely to exhibit significant outperformance measures. The results show that ETFs with either high expense ratios or recent inception dates are more likely to display significant outperformance. The process for replication evolves and becomes more refined over time, hence such outperformance for recently incepted ETFs. Table 2.9 shows that the outperformance is primarily due to index TEs being present; in other words, these funds outperform their underlying indices. The expense ratio result is in line with Elton et al. (2002) and Harper et al. (2006), who find that more expensive ETFs tend to produce greater returns but the difference dissipates once the increased market frictions are accounted for. The economic interpretation here is that ETFs with more profitable replication strategies charge more. In addition, larger ETFs have a greater tendency to display significant Sharpe ratios premiums than smaller ETFs. This could be due to larger ETFs finding rebalancing more difficult as it takes greater resources for less nimble ETFs to rebalance accurately.

Table 2.10 shows the top 10 funds under each performance measure, compiled and ranked using mean daily outperformance figures. The ETFs in the top 10 for Sharpe and Sortino ratios across various performance measures highlight the interdependency between these calculations. The distinction between these standard deviation-based ratios and the

Treynor ratio, which utilises the CAPM derived β, or co-movement between the market and ETF price, as a risk proxy, is apparent when analysing the cross-measure top ten ranking composition.

PCY US and LQD US are tickers of particular interest as they appear in the top three NAV and index outperformers under both the Sharpe and Sortino ratio measures. PCY US is the ticker symbol for the PowerShares Emerging Market Sovereign Debt Portfolio, which is based on the DB Emerging Market USD Liquid Balanced Index. Its portfolio is comprised of U.S. dollar-denominated government bonds issued by, at present, in 64 emerging market countries.[17] It is one of the more recent ETFs in the data set, being incepted in October 2007. It follows a full replication strategy with an expense ratio of 0.50%. The volatility and diversity of the underlyings creates replication difficulties, particularly for a recently incepted ETF which can lead to such substantial outperformance.

LQD US is the ticker symbol for iBoxx $ Investment Grade Corporate Bond Fund, which tracks the iBoxx $ Liquid Investment Grade Index. Its portfolio is comprised of liquid, U.S. dollar-denominated, investment-grade corporate bonds for sale in the U.S.. It is a cross-sectoral fund with over 34% currently invested in financial services.[18] Its inception date of the July 26, 2002 is older than the data set median. It also follows a full replication strategy with an expense ratio of 0.15%. Combining the volatile financial services sector with corporate credit rating re-evaluations due to increased market fear and slow reacting debt markets gives rise to substantial outperformance. These ETFs provide an insight into the attribute mix of ETF whose prices *substantially* outperform their NAVs and underlying indices.

DBS US and SLV US are tickers in the top three market benchmark outperformers across both Sharpe and Sortino ratios. DBS is the Powershares DB Silver ETF, with SLV the ticker symbol for the iShares Silver Trust. Both funds provide exposure to the market price of silver which *substantially* outperforms the market over the period. Demand from private investors in Asia is a mooted reason for the increase in the price of silver.

2.6 Conclusion

This study seeks to identify ETFs that outperform their calculated NAVs, underlying indices, and/or the overall market. Extending the existing ETF literature, an innovative generalised stepwise procedure is used to control for data snooping bias. We argue in this paper that controlling for data snooping bias is important in order to obtain robust economic conclusions on ETF performance. This paper is the first amongst the ETF literature to take this approach. The balanced stepdown procedure of Romano and Wolf (2010) is applied, serving as an improvement over more conservative single step approaches, such as common techniques like the reality check bootstrap test of White (2000) and the superior

[17] PowerShares PCY emerging markets sovereign debt portfolio fund holdings: http://www.invescopowershares.com/products/holdings.aspx?ticker=PCY (accessed 10/30/12).

[18] iShares iBoxx dollar investment grade corporate bond ETF factsheet: http://us.ishares.com/content/stream.jsp?url=/content/en_us/repository/resource/fact_sheet/lqd.pdf (accessed 10/30/12).

predictive ability test of Hansen (2005). generalised procedures offer greater power to reject false null hypotheses, with the balanced stepdown procedure additionally offering equal treatment in the identification of outperformance. Given the geographical and sector coverage of the extensive ETF database considered, we provide first-stage guidance to investors of where inefficiencies may be and, accordingly, where ETFs may provide some investment advantages. The main item of note from the implementation is that, when performance is analysed on a non-risk-adjusted basis only, no funds in the sample are identified as displaying any measure of outperformance. It is only the risk-adjusted performance measures that give statistically significant outperformance results and so the insights from these results dominate the commentary. The three key takeaways from the study are, firstly, a high proportion of optimised replication, debt asset class, and global/international ETFs exhibit risk-adjusted premiums, highlighting redemption in kind inefficiencies. Secondly, cross-sector and sectoral funds display broadly the same percentage of outperformance. Finally, high expense ratio and recent inception date ETFs are more likely to exhibit index outperformance, which may be of interest to investors seeking to outperform benchmarks.

Our study is the first to test the effect of replication type on performance. We find that 50% of optimised replication ETFs register significant Sharpe ratio premiums. This finding may, in part, be due to trading restrictions and by less than optimal replication strategies. We are also the first to examine asset class focus. We find that 78% of debt-focused ETFs exhibit significant Sharpe ratio premiums, which is well above the average and gives an indication that debt-focused ETFs are more likely to outperform their NAV than other asset classes. The performance of sectoral ETFs on the other hand has been addressed previously. In this work, energy, precious metals and real estate are industries that beat the market on a risk-adjusted basis. Further to this, precious metal-focused funds Powershares DB Silver and the iShares Silver Trust *substantially* outperform the market, exhibiting large mean daily outperformance levels. Precious metals became a safe haven for investors due to poor performance in equities over the turbulent 2008-2012 period, with the energy sector being buoyed by increased manufacturing demand from China. [19] The financial services sector, in contrast, registers no market beating funds, primarily due to the credit crisis of 2007-2009 and its legacy.

Global (63%) and international (79%) ETFs, show premium Sharpe ratio outperformance compared to U.S. funds (10%). These findings are in line with and Jares and Lavin (2004) and Engle and Sarkar (2006), who also observe premiums among a high percentage of foreign ETFs. Elton et al. (2002) and Ackert and Tian (2008) record a low proportion of U.S. focused funds displaying premiums. A lack of synchronisation between NAV calculations and underlying market closes is an oft-cited reason for the presence of premiums in funds focused over multiple countries/time zones. Furthermore, liquidity, latency advantages, and reduced market frictions allow for easier exploitation of deviations among U.S. focused funds. ETFs exhibiting high expense ratios or recent inception dates have

[19] Given the strength of the flight to safe haven assets observed in this 2008-2012 period investors are urged to thread with caution when utilising this result for portfolio selection decisions out-of-sample.

a greater tendency to outperform their index. This expense ratio result is in line with Elton et al. (2002) and Harper et al. (2006), who find that more expensive ETFs tend to produce greater returns but the difference dissipates once the increased market frictions are accounted for. This paper succeeds in increasing the understanding of ETF performance alongside providing investors with first-stage guidance in identifying ETFs suitable for their portfolios.

Table 2.1: Data set properties

Industry Focus	Count
Cross Sector	193
Energy	18
Technology Sector	14
Financial Services	12
Health & Biotechnology	11
Real Estate Sector	10
Utility Sector	7
Precious Metals Sector	7
Environmentally Friendly	4
Internet/Telecommunications	4
Leisure Industry Sector	2
Food/Beverage Sector	1
No Industry Focus Given	5

Geographic Focus	Count
United States	188
International	34
Global	27
China	5
European Region	3
Japan	3
Asian Pacific Region ex Japan	2
Latin American Region	2
Other	24

Asset Allocation	Count	Derivative Replication	Count
Equity	263	Full	145
Commodity	13	optimised	62
Debt	9	Unknown	46
Asset Allocation	3	Derivative	35

Count of ETFs in data set split by various attributes. "No Industry Focus Given" is used to denote sector ETFs where no industry focus has been provided by Bloomberg.

Table 2.2: Outperformance measures

Outperformance Measure	**Assigned Name**
ETF price log return− ETF NAV log return	Premium
ETF price log return−Underlying index's log return	Index TE
ETF price log return−S&P 500 log return	Mkt TE
ETF price log return Sharpe ratio− ETF NAV log return Sharpe ratio	Premium SR
ETF price log return Sharpe ratio−Underlying index's log return Sharpe ratio	Index TE SR
ETF price log return Sharpe ratio−S&P 500 log return Sharpe ratio	Mkt TE SR
ETF price log return Sortino ratio−ETF NAV log return Sortino ratio	Premium SorR
ETF price log return Sortino ratio−Underlying index's log return Sortino ratio	Index TE SorR
ETF price log return Sortino ratio−S&P 500 log return Sortino ratio	Mkt TE SorR
ETF price log return Treynor ratio−ETF NAV log return Treynor ratio	Premium TR
ETF price log return Treynor ratio−Underlying index's log return Treynor ratio	Index TE TR

Table 2.3: Significant sample summary statistics

	Mean	Std. Dev.	Max (ETF Ticker)	Min (ETF Ticker)
Premium			N/A	
Index TE			N/A	
Mkt TE			N/A	
Premium SR	0.02567	0.04385	0.32094 (PCY)	0.00101 (FIW)
Index TE SR	0.02859	0.05158	0.39317 (PCY)	0.00190 (DBC)
Mkt TE SR	0.03228	0.01447	0.06579 (AGG)	0.01090 (PLW)
Premium SorR	0.19738	0.28069	1.88967 (DBS)	0.01190 (SLY)
Index TE SorR	0.21858	0.34545	2.77124 (QLD)	0.01060 (IJH)
Mkt TE SorR	0.25455	0.11153	0.51299 (PCY)	0.07473 (VXF)
Premium TR	0.00751	0.01465	0.08015 (GXC)	0.00002 (RWM)
Index TE TR	0.00861	0.01130	0.06282 (AGG)	0.00001 (IJJ)

Mean (Column 2) refers to the average daily outperformance levels across the 2008-2012 period. Max and Min (Columns 4 & 5) identify those ETF tickers which display the highest and lowest aggregated daily outperformance level. All funds are U.S.- based with the Bloomberg ticker appendage "U.S." being omitted for brevity.

Table 2.6: ETFs outperformance by replication type

	Full	optimised	Derivative	Unknown
	% (Count)	*% (Count)*	*% (Count)*	*% (Count)*
Premium	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	27% (39)	50% (31)	26% (12)	9% (3)
Index TE SR	21% (31)	19% (12)	17% (8)	26% (9)
Mkt TE SR	21% (30)	11% (7)	22% (10)	26% (9)
Premium SorR	25% (36)	40% (25)	28% (13)	3% (1)
Index TE SorR	24% (35)	35% (22)	22% (10)	26% (9)
Mkt TE SorR	37% (54)	40% (25)	37% (17)	26% (9)
Premium TR	15% (22)	55% (34)	46% (16)	39% (18)
Index TE TR	23% (33)	52% (32)	57% (20)	37% (17)

Percentage of ETFs in each replication strategy which have specific outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in parenetheses gives the ETF count in each group.

Table 2.4: ETF outperformance by geographic focus

	United States	International	Global	China	Europe	Japan	Asia Pac	Latin America	Other
	% (Cnt)	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*
Premium	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	10% (18)	79% (27)	63% (17)	60% (3)	67% (2)	0% (0)	0% (0)	100% (2)	67% (16)
Index TE SR	11% (20)	44% (15)	63% (17)	20% (1)	0% (0)	0% (0)	0% (0)	50% (1)	25% (6)
Mkt TE SR	27% (51)	0% (0)	7% (2)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	13% (3)
Premium SorR	10% (18)	71% (24)	56% (15)	60% (3)	67% (2)	0% (0)	0% (0)	50% (1)	50% (12)
Index TE SorR	15% (28)	47% (16)	70% (19)	40% (2)	0% (0)	0% (0)	0% (0)	50% (1)	38% (10)
Mkt TE SorR	41% (78)	18% (6)	19% (5)	40% (2)	0% (0)	0% (0)	50% (1)	100% (2)	42% (11)
Premium TR	13% (24)	68% (23)	48% (13)	40% (2)	100% (3)	0% (0)	50% (1)	100% (2)	92% (22)
Index TE TR	14% (27)	82% (28)	74% (20)	20% (1)	100% (3)	33% (1)	0% (0)	100% (2)	83% (20)

Percentage of ETFs in each geographic focus category that display specific outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in parentheses gives the ETF count in each group. "Cnt" is an abbreviation of Count and "Asia Pac" is an abbreviation of Asia Pacific excluding Japan, both are used for brevity.

Table 2.5: ETF outperformance by industry

	Energy Sector	Tech Sector	Financial Services	Health & Biotech	Real Estate	Utility Sector	Precious Metals	Environ Friendly	Internet/ Telecoms	Leisure Industry	Cross Sector	N.A.
	% (Cnt)	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*	*% (Cnt)*
Premium	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	28% (5)	7% (1)	17% (2)	9% (1)	30% (3)	43% (3)	14% (1)	25% (1)	25% (1)	0% (0)	33% (64)	60% (3)
Index TE SR	28% (5)	14% (2)	17% (2)	9% (1)	0% (0)	43% (3)	14% (1)	50% (2)	25% (1)	0% (0)	21% (40)	60% (3)
Mkt TE SR	56% (10)	14% (2)	0% (0)	0% (0)	50% (5)	0% (0)	71% (5)	0% (0)	25% (1)	100% (2)	16% (30)	20% (1)
Premium SorR	17% (3)	7% (1)	17% (2)	18% (2)	10% (1)	43% (3)	14% (1)	25% (1)	25% (1)	50% (1)	29% (56)	60% (3)
Index TE SorR	28% (5)	14% (2)	17% (2)	18% (2)	10% (1)	57% (4)	0% (0)	75% (3)	25% (1)	50% (1)	27% (52)	60% (3)
Mkt TE SorR	61% (11)	50% (7)	0% (0)	36% (4)	60% (6)	14% (1)	100% (7)	0% (0)	50% (2)	100% (2)	33% (64)	20% (1)
Premium TR	6% (1)	14% (2)	17% (2)	18% (2)	40% (4)	43% (3)	0% (0)	25% (1)	25% (1)	0% (0)	38% (73)	20% (1)
Index TE TR	22% (4)	14% (2)	17% (2)	9% (1)	50% (5)	43% (3)	0% (0)	25% (1)	25% (1)	0% (0)	42% (82)	20% (1)

Percentage of ETFs in each industry focus category that displays specific outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in parentheses gives the ETF count. "N.A." is used to denote sector ETFs where no industry focus has been provided by Bloomberg. "Cnt" is an abbreviation of Count and "Environ" is an abbreviation of Environmentally, both are used for brevity.

Table 2.7: ETFs displaying specific outperformance by asset class

	Equity	Commodity	Debt	Asset Allocation
	% (Count)	*% (Count)*	*% (Count)*	*% (Count)*
Premium	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	28% (73)	23% (3)	78% (7)	67% (2)
Index TE SR	19% (51)	31% (4)	22% (2)	100% (3)
Mkt TE SR	19% (51)	38% (5)	0% (0)	0% (0)
Premium SorR	25% (65)	8% (1)	89% (8)	33% (1)
Index TE SorR	25% (67)	15% (2)	44% (4)	100% (3)
Mkt TE SorR	38% (99)	46% (6)	0% (0)	0% (0)
Premium TR	33% (88)	0% (0)	11% (1)	33% (1)
Index TE TR	38% (99)	0% (0)	22% (2)	33% (1)

Percentage of ETFs in each asset class which specific outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in parentheses gives the ETF count in each group.

Table 2.8: Outperformance by asset/ER/inception date

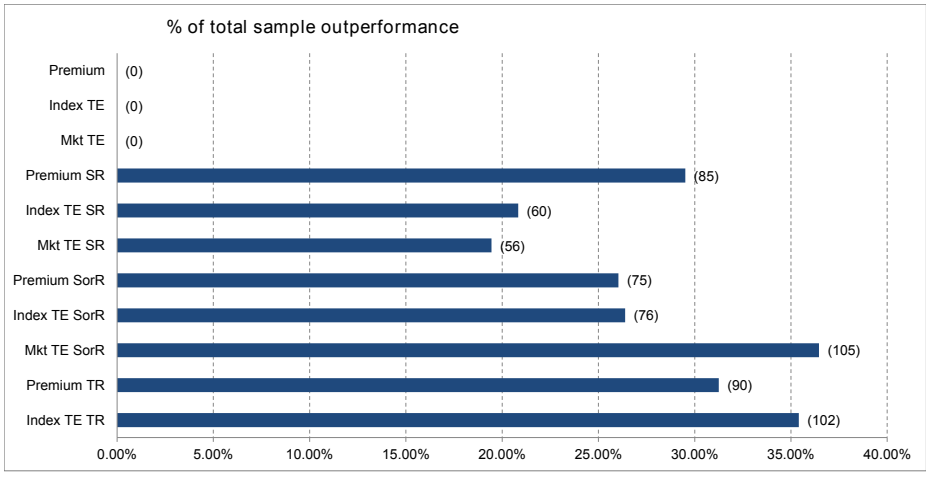
	Assets ($M)	Expense Ratio	Inception Date
Data Set			
Mean	2965.02	0.52%	
Median	421.89	0.51%	15/09/2005
Outperforming ETFs			
Mean	2774.50	0.56%	
Median	429.92	0.52%	01/02/2006
# $\geq$ Data set median (%)	107 (51%)	132 (63%)	121 (58%)
# $<$ Data set median (%)	103 (49%)	78 (37%)	89 (42%)

Table 2.9: ETFs outperformance by total assets/expense ratio/inception date

	Total Assets		Expense Ratio		Inception Date	
	$\geq$ *$421.89m*	< *$421.89m*	$\geq$ *0.51%*	< *0.51%*	$\geq$ *15-Sep-05*	< *15-Sep-05*
	% (Count)	*% (Count)*	*% (Count)*	*% (Count)*	*% (Count)*	*% (Count)*
Premium	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Index TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Mkt TE	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Premium SR	37% (53)	22% (32)	35% (50)	24% (35)	36% (52)	23% (33)
Index TE SR	34% (49)	18% (26)	27% (39)	25% (36)	28% (41)	24% (34)
Mkt TE SR	8% (11)	6% (8)	10% (14)	3% (5)	9% (13)	4% (6)
Premium SorR	8% (11)	15% (21)	19% (28)	3% (4)	18% (26)	4% (6)
Index TE SorR	29% (42)	24% (34)	33% (47)	20% (29)	33% (48)	20% (28)
Mkt TE SorR	23% (33)	19% (27)	26% (38)	15% (22)	30% (43)	12% (17)
Premium TR	37% (53)	26% (37)	40% (57)	23% (33)	34% (50)	28% (40)
Index TE TR	36% (52)	35% (50)	46% (66)	25% (36)	43% (62)	28% (40)

Percentage of ETFs with outperformance measures under the balanced stepdown procedure of Romano and Wolf (2010). The figure in parentheses gives the ETF count.

Figure 2.1: % of ETFs with specific outperformance measures



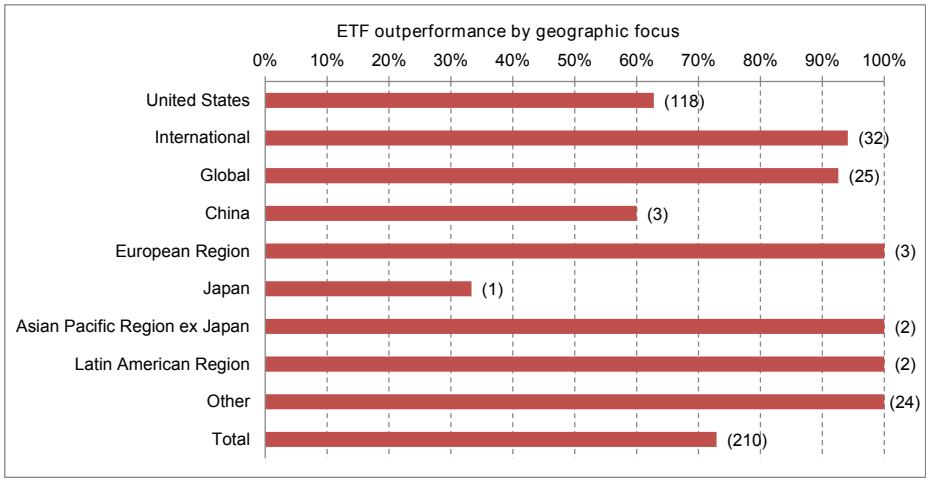
Percentage of ETFs with specific outperformance measures deemed significant under the balanced stepdown procedure of Romano and Wolf (2010). The figure in parentheses gives the fund count in each group.

Table 2.10: Top 10 ETFs by mean daily outperformance

Rank	Premium SR	Premium SorR	Premium TR	Index TE SR	Index TE SorR	Index TE TR	Mkt TE SR	Mkt TE SorR
1	PCY (0.32094)	PCY (1.88967)	GXC (0.08015)	PCY (0.39317)	PCY (2.77124)	AGG (0.06282)	QLD (0.06579)	DBS (0.51299)
2	LQD (0.18214)	LQD (1.06739)	EWM (0.0721)	LQD (0.13133)	PZA (1.03797)	EWM (0.05754)	DBS (0.05919)	SLV (0.50375)
3	HYG (0.16968)	HYG (1.04502)	AGG (0.06213)	DDM (0.05353)	LQD (0.82621)	HYG (0.04979)	SLV (0.05902)	U.S.D (0.49500)
4	MUB (0.11583)	MUB (0.95198)	EPP (0.05059)	RSU (0.04903)	MUB (0.79061)	EWH (0.04763)	MVV (0.05857)	DIG (0.49411)
5	AGG (0.08780)	EMB (0.43926)	EWH (0.04002)	SSO (0.04833)	DDM (0.43859)	EWT (0.03502)	UWM (0.05600)	QLD (0.49252)
6	EMB (0.05460)	AGG (0.43746)	EWT (0.02934)	MVV (0.04616)	RSU (0.41887)	AIA (0.02783)	DIG (0.05532)	UWM (0.48788)
7	IXJ (0.03775)	PZA (0.40871)	AIA (0.02329)	IXJ (0.03992)	SSO (0.38298)	GMF (0.02546)	SAA (0.05512)	MVV (0.48707)
8	DGS (0.03704)	DGS (0.32111)	EWS (0.02185)	KXI (0.03920)	UUP (0.34833)	ITF (0.02038)	U.S.D (0.05172)	SAA (0.46096)
9	UUP (0.03694)	PBP (0.27846)	DBV (0.01938)	SAA (0.03878)	MVV (0.33571)	EWS (0.02016)	PXI (0.04854)	PXI (0.44491)
10	DEM (0.03583)	IXJ (0.27723)	IYM (0.01879)	GAF (0.03800)	JXI (0.33035)	EEV (0.01875)	PXE (0.04696)	PXE (0.43693)

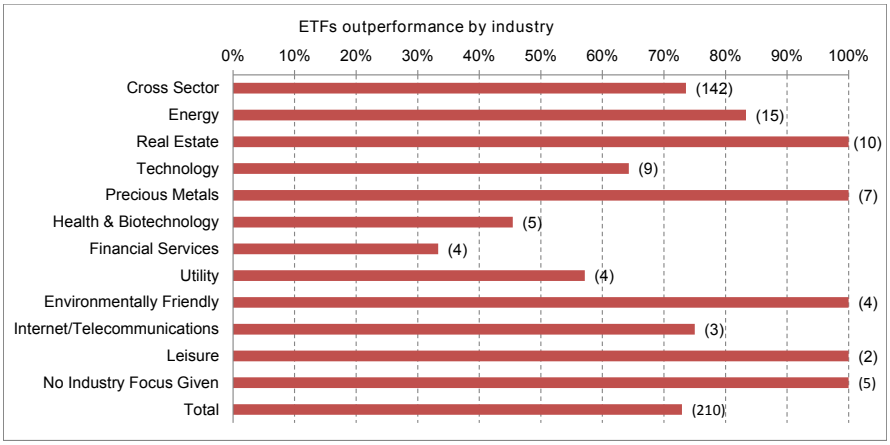
Significant outperforming ETFs under the balanced stepdown procedure of Romano and Wolf (2010) are ranked by the size of their mean daily outperformance measures. The outperformance measure figure is given in parentheses. All funds are U.S. based with the Bloomberg ticker appendage "U.S." being omitted for brevity.

Figure 2.2: % of ETFs displaying outperformance by geographic focus



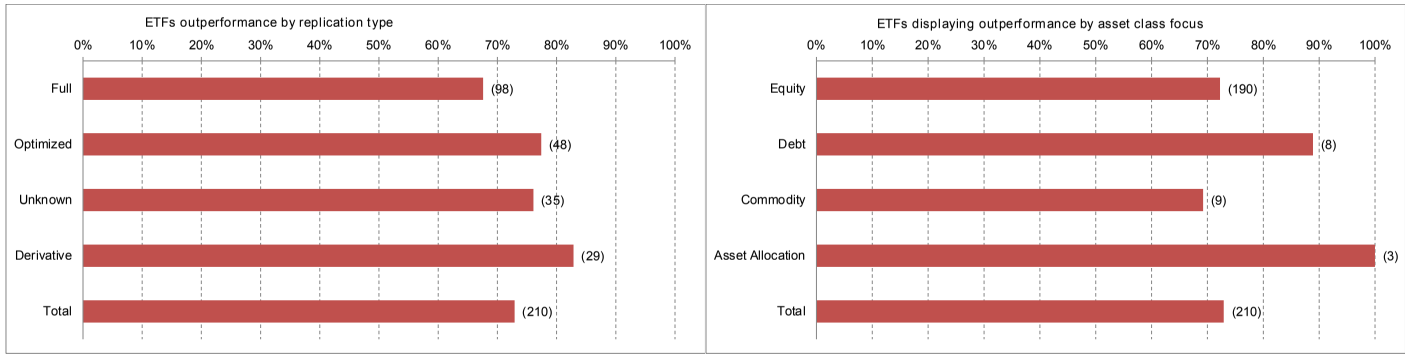
Percentage of ETFs in each geographic focus category with at least one significant outperformance measure under the balanced stepdown procedure of Romano and Wolf (2010). The figure in parentheses gives the ETF count in each group.

Figure 2.3: % of ETFs displaying outperformance by industry focus



Percentage of ETFs in each industry focus category which display at least one significant outperformance measure under the balanced stepdown procedure of Romano and Wolf (2010). The figure in parentheses gives the ETF count in each group.

Figure 2.4: % of ETFs displaying outperformance by replication type/asset class focus



Percentage of ETFs in each asset class focus category and also in each replication type category which display at least one significant outperformance measure under the balanced stepdown procedure of Romano and Wolf (2010). The figure in parentheses gives the ETF count in each group.

Chapter 3

An analysis of implied volatility jump dynamics: Novel functional data representation in crude oil markets

3.1 Introduction

Oil futures are the most actively traded commodity derivatives. An average of one million light sweet crude oil futures and option contracts are traded every day according to the CME group.[1] The past 10 years have seen elevated levels of price volatility in these markets. Strong economic pressures have been observed on both the demand side and the supply side, during the global financial crisis and the Arab Spring respectively. Increased price volatility in oil markets causes profound economic management and socio-political issues, not only impacting those participants who invest directly in commodities but also the consumers of refined oil products. We use a functional data analysis (FDA) approach to examine implied volatility, jump risk, and pricing dynamics within crude oil markets. FDA provides a framework to uncover the continuous processes underlying a data set. The process of interest in this study is that of the implied volatility curve. The FDA methodology has many advantages; it accurately captures implied volatility dynamics, there is no assumed parametric structure, it is computationally efficient, and the process can be evaluated on an arbitrarily fine grid.[2] This facilitates the consistent comparison of identical option contracts through time. Implied volatility is of interest as it is a transformation of the option price and a key parameter in many asset pricing and regulatory capital calculations. Implied volatility also contains informational content as shown by Corrado and Miller 2006, Taylor et al. 2010, Muzzioli 2010, and Garvey and Gallagher 2012. Furthermore, Yan (2011) proposes the use of implied volatility slope information to estimate jump risk. He demonstrates both directly and indirectly, the applicability of at-the-money implied

[1] http://www.cmegroup.com/trading/energy/files/en-153_wti_brochure_sr.pdf accessed on 11/11/2013.

[2] The FDA advantages listed here are outlined in Ramsay and Silverman (2005).

volatility slope as a proxy for the average jump amplitude in equity markets. We seek to answer three related questions pertaining to crude oil markets. Firstly, what is the link between the shape of the implied volatility smile and underlying economic events in crude oil markets? Secondly, does the implied volatility slope contain information useful in specifying the average jump amplitude for crude oil options, in a similar manner to what Yan (2011) has shown to be the case for stock returns?[3] Thirdly, can information contained in the implied volatility smile slope be exploited to improve portfolio hedging techniques?

Traditional geometric Brownian motion (GBM) based models, such as Black and Scholes' (1973) diffusion model ("Black-Scholes" henceforth), do not capture price jumps, which are movements that become more prevalent during periods of increased market turbulence. There is a large body of literature demonstrating the importance of incorporating jumps into models seeking to capture risk premia and economic shocks. Cont and Tankov (2004) argue that discontinuous shifts are the most important element of pricing in crude oil markets, and propose the use of a Lévy process to model such movements.[4] Askari and Krichene (2008) find that 2002-2006 WTI oil price dynamics are dominated by jumps; the variance due the fitted model's diffusion component is high and significant, but the variance due to the jump component is even higher. In line with Yan (2011), we adopt the Merton (1976) jump diffusion framework ("Merton model" henceforth), which augments the Black-Scholes diffusion model with a jump process in order to model continuous price innovation and discontinuous price shock movements simultaneously. The model has two elements: a Black-Scholes drift for capturing regular price movements and a jump component for capturing large irregular and infrequent price shifts. The continuous diffusion follows a GBM model with constant drift and volatility, while the discontinuous jumps are modelled by a Poisson process. The relationship between the volatility smile shape and the parameters of the Merton jump diffusion model is assessed both theoretically and empirically. As in Yan (2011), this one-dimensional Brownian motion and Poisson process model is adopted for tractability. Using a 2007-2013 sample period, strong evidence is found of converse jump dynamics during periods of demand and supply side weakness. Furthermore, using FDA to systematically analyse underlying economic forces highlights periods of economic weakness in advance of their occurrence.

The specification of the correct jump parameter level is necessary to accurately capture implied volatility curve skewness and kurtosis, according to Borensztein and Dooley (1987), Jorion (1988), and Bates (1996). The jump component in Askari and Krichene (2008) for instance, displays high intensity and variance, with a negative mean jump size being associated with negative skewness in the empirical distribution. Nomikos and Soldatos (2010) find that the presence of jumps in the related power options market, generates

[3]Chang et al. (2013) emphasise the importance of crude oil markets by outlining a strong link with stock market movements.

[4]Bakshi et al. (1997) and Trolle and Schwartz (2009) also advocate the incorporation of jumps in the accurate pricing of short-term derivatives, with Wilmot and Mason (2013) emphasising the importance of jumps when studying daily data.

implied volatility skews, which again depend on the sign of the mean jump size. Our study outlines the effect of Merton model parameter specification on implied volatility curve shape and build on this relationship to specify the level of average jump size using a simple FDA-derived proxy.

A number of applications in the bio-mechanical literature have included FDA; however, it has only recently been exploited for financial analysis. Benko et al. (2009) propose a new functional principal component analysis (fPCA) technique to study similarities in the implied volatility dynamics for different maturities using both one- and three-month option maturities on the German-Swiss exchange. In doing so, they find that FDA accurately captures the implied volatility dynamics. Muller et al. (2011) propose a functional volatility process to model volatility trajectories for high-frequency observations in financial markets. Their model shows patterns in volatility and by combining it with prediction techniques and functional regression, it can be used to predict future volatility. Our research is unique as it utilises FDA representation of implied volatility curves to analyse changes in economic forces over time, specify Merton model jump parameters, and empirically demonstrate portfolio hedging benefits.

The relationship between jump amplitude and implied volatility skewness is studied in detail. This is achieved by combining FDA techniques with the Merton model to extract implied jump size probability and direction. The significance here is that the continuous implied volatility function can be differentiated to obtain the slope and other higher order derivatives. We show how the slope levels at various moneyness points provide great insight into the demand and supply forces observed. A strong economic link is also made between the average jump amplitude level and these demand and supply side forces. This leads to the FDA obtained at-the-money (ATM) slope being utilised as a novel proxy for the average jump amplitude value in specifying the Merton model. Further contributions of our paper relate to the employment of FDA obtained Merton model parameters for portfolio hedging. We compare the calculated results against the standard Black-Scholes delta hedging strategy. The Merton delta hedging strategy outperforms the Black-Scholes delta hedging strategy by 8% in terms of implementation cost, over the entire sample. Breaking the sample down into periods split by predominantly positive and predominantly negative implied volatility slopes, we see that the Merton strategy outperforms the Black-Scholes when jump direction is positive and broadly matches its performance when jump direction is negative.

The rest of the paper is organised as follows. Section 5.3 introduces the FDA methodology, analyses the dynamics of varying parameters in the Merton model, and presents some stylised facts about crude oil options. Section 4.4 details the data set utilised, while Section 3.4 presents a discussion of the implied volatility curve shape in terms of demand and supply side weaknesses observed in the sample period. The optimised delta hedging results are also reported in Section 3.4, with concluding remarks given in Section 5.5.

3.2 Methodology

3.2.1 Implied volatility curve shape

The shape of the implied volatility curve is formed through the interaction of economic forces, namely demand and supply. The inherent fear in equity markets is that of a price crash, similar to the losses observed on days such as Black Monday 1987. When fear of a price crash is prevalent in the market, it tends to lower OTM implied volatility while raising ITM implied volatility.[5] Bates (1991) shows an S&P 500 volatility curve for call options that demonstrate a negative (reverse) skew, whereby in-the-money (ITM) call options exhibit a higher implied volatility than their out-of-the-money (OTM) counterparts in periods preceding such crashes. It is associated with the intuition that investors are willing to pay more for downside protection. On the contrary however, when fear of a price spike is prevalent in the market, it tends to lower ITM implied volatility while raising OTM volatility. Evidence of this dynamic is put forward for commodities in Askari and Krichene (2008) and Liu and Tang (2011). They investigate crude oil markets and find a distribution that is positively (forward) skewed.[6] It is interpreted from an economic perspective that commodity market participants assign a higher relative value to OTM options in comparison to their ITM counterparts due to a fear of price spikes. This interesting dynamic is a primary reason why crude oil options are chosen for this study. Commodity investors are less worried about downward drops in prices than upward jumps. Under these circumstances, the intuition is that investors are willing to pay more for upside protection.

Figure 3.1 shows the typical skew pattern in the crude oil implied volatility curve for a WTI CL02 call option according to the forward skew identified by Askari and Krichene (2008) and Liu and Tang (2011). The curve is steeper at the OTM point to indicate fear of a price spike. A higher absolute slope value is recorded at the OTM point versus that of the ITM. Furthermore, the volatility used to price ITM options is lower than the volatility used to price OTM options. This shape corresponds to an implied distribution with heavier right tail and a less heavy left tail than that of the Black-Scholes assumed lognormal distribution. There is a higher price and hence higher volatility attached to OTM options to protect against the expectation of positive market jumps.

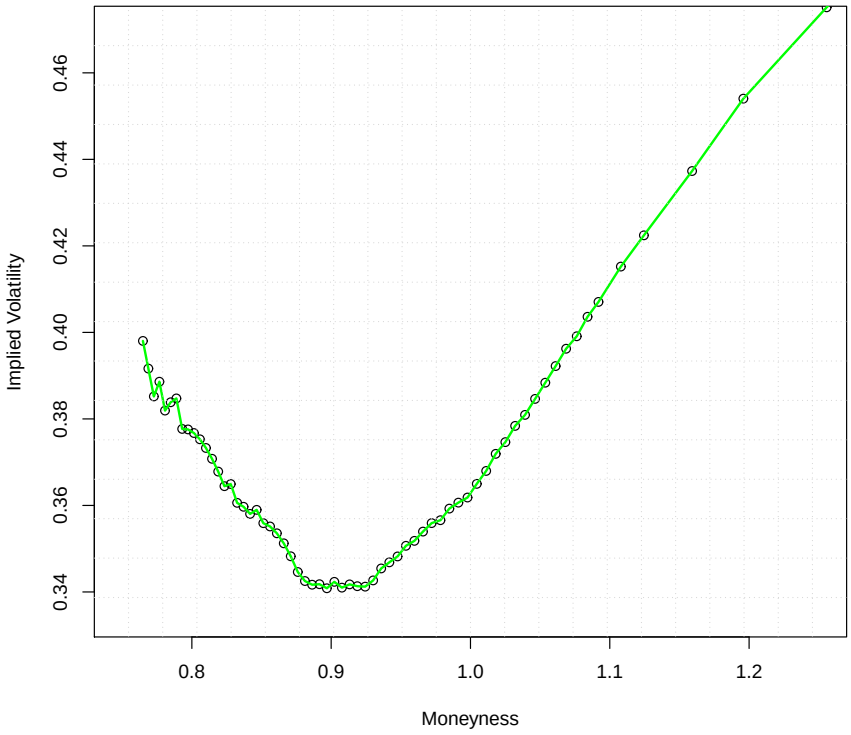
3.2.2 Functional data representation

Our analysis extends the prevailing literature on crude oil implied jump dynamics by combining FDA techniques with the Merton model to extract implied jump size probability and direction. FDA uncovers the smooth process underlying the data. This sets it apart

[5] The commentary presented here is for call options, the opposite effect is true for put options.

[6] The presence of a skew is also in line with studies by Mandelbrot (1963), Fama (1965), and Clark (1973).

Figure 3.1: Typical crude oil implied volatility curve (11th June 2010)



A typical crude oil option implied volatility graph (11 June 2010) for CL02 WTI crude oil option (1-2 month maturity).

from regular interpolation techniques which simply seek to find the best fit to the possibly noisy data set. This true function is represented in an infinite dimensional space over a continuum of values. A continuum is generally defined in terms of time or space; however, in this paper the functions are defined in the moneyness domain, as an option's implied volatility is related to how ITM [i.e., K/F (strike/forward rate)] it is. The FDA methodology has many advantages; it accurately captures volatility dynamics, there is no assumed parametric structure, it is computationally efficient, and the process can be evaluated on an arbitrarily fine grid. This FDA representation allows us to consistently identify the ATM (and ITM and OTM) implied volatility level for each day. It serves as an improvement over Yan (2011)'s near the money implied volatility calculation that only encompasses information obtained from -0.5 delta puts and 0.5 delta calls.

Our goal is to interpret the daily discrete option volatility data, $x(m_k)$, as a functional data object or, more simply, as a function, denoted $\tilde{x}(m_k)$.[7] When constructing a function, a vector of n basis function, denoted $\phi_{1,...,n}$, must first be specified. Functional structures are approximated as a weighted linear combination of these bases:

$$\tilde{x}(m) = c_1\phi_1(m) + c_2\phi_2(m) + ... + c_n\phi_n(m),$$

where $c_1, ..., c_n$ represent the parameters of the expansion's coefficients.

As in Ramsay and Silvermann (2005), the coefficients c_j are chosen in order to minimise:

$$SSE(c_1, ..., c_n) \equiv \sum_{k=1}^{N} \left[x(m_k) - \sum_{j=1}^{n} c_j\phi_j(m_k) \right]^2 \qquad (1).$$

The decision of which basis system to specify is driven by the underlying data's known characteristics. For instance, when modelling periodic data, a Fourier basis expansion, comprised of successive sine/cosine terms, is most commonly applied. However, an implied volatility process does not exhibit strong cyclical variation, so B-spline functions are chosen for the basis function system. B-splines are essentially a number of polynomials joined together smoothly at fixed points called knots. The number and positioning of the N knots: $m_k : m_1 \leq ... \leq m_N$, are derived from knowledge of the complexity of the underlying process over particular ranges. The range of the various sub-intervals, $[m_k, m_{k+1}]$, are defined through the placement of these knots. Within each sub-interval, the spline is simply a polynomial of order n. The order is calculated as:

$$order = 1 + degree\ of\ the\ polynomial.$$

B-spline representation is useful as it has a number of strengths. At any one point along the curve, it simplifies to a polynomial that can be easily evaluated. Adjusting the order of the spline allows for the estimation of derivatives of any degree. In this paper, a fourth

[7] The standard notation utilised, $x(t)$, signifies a dependence on time; however, here the domain is that of the moneyness (K/F) levels, hence the use of $x(m)$.

order basis, or cubic polynomial is specified. Modelling the process as a cubic polynomial provides a good balance as it retains the function's continuous property up to the second derivative, while simultaneously smoothing the noise within the daily data.[8] The knots are placed at the discrete quoted option moneyness levels that are available from the data set, with polynomials describing the moneyness interval between the knots.

Given that a smooth function underlies the observed implied volatility curve a smoothing penalty is applied to remove noise/wiggles in the data. This can be caused by liquidity issues, misquotes or other data irregularities masking the true function. The effect of such non-trading and noise issues on the volatility function has been highlighted by Bannouh et al. (2013). Without the use of a smoothing penalty the uncovered function may simply converge to a fitted spline interpolant of the data. In line with Ramsay and Silvermann (2005) a limitation is placed on the variation of the curvature. The total curvature of the process is found through integrating its squared second derivative:[9]

$$R(\tilde{x}) \equiv \int \left(\frac{d^2}{dm^2} \tilde{x}(m) \right)^2 dm.$$

This is also called the roughness of the function.

In an extension of (1), the coefficients characterising the smoothed curve are found using the penalised sum of squared errors:

$$PENSSE(c_1, ..., c_n) \equiv \sum_{k=1}^{N} \left[x(m_k) - \sum_{j=1}^{n} c_j \phi_j(m_k) \right]^2 + \lambda^* \int \left(\frac{d^2}{dm^2} \tilde{x}(m) \right)^2 dm.$$

As λ^* increases, more weight is placed on the roughness penalty and the uncovered function converges towards a straight line, possibly missing some of the process' dynamics.[10] As λ^* decreases, less weight is placed on the roughness penalty and only data fitting matters in uncovering the function.

In order to balance the competing goals of retaining features and removing noise from the data, an optimal smoothing level, λ^*, must be selected. Inherent knowledge of variation in the underlying process is useful here and can be used in conjunction various data-driven techniques such as information criteria and cross validation. Cross validation is based on the principal of removing one observation from the sample and using this sub-sample to see how well the removed observation can be predicted. We apply generalised cross validation (GCV), proposed by Craven and Wahba (1979). Without smoothing, noise in the relatively small number of discrete values available can distort the results for that range. This is particularly evident at extreme moneyness levels as prices may be contorted due to illiquidity.

[8] This is required as we examine the curvature of the implied volatility function.

[9] The second derivative is also denoted as $D^2x(m)$ in the literature.

[10] The symbol λ^* is used here to distinguish this smoothing weight for calculating coefficients under penalised sum of squared errors, from the jump intensity λ parameter specified in the Merton model.

FDA allows a systematic analysis of implied volatility curves over time. It provides a framework to obtain the slope and curvature at any point along the implied volatility curve. Therefore, the slope and curvature levels are evaluated consistently at three points along the daily implied volatility curve, namely, ATM, 10% ITM, and 10% OTM. The evolution of these levels over time is analysed in order to further understand how demand and supply side weaknesses alter its shape.

3.2.3 Merton model

In line we Yan (2011) we adopt the one-dimensional Brownian motion and one-dimensional Poisson process of Merton (1976) to model discontinuous price moves. Yan (2011) note that it can be extended to incorporate multi-dimensional Brownian motions and Poisson processes. Under the Merton (1976) jump diffusion (JD) framework, Poisson jumps are combined with a continuous Black and Scholes (1973) diffusion model. A call option is priced as:

$$j(\lambda', T)c_n(F_T, X, T, r_n, q, \sigma_n),$$

$$\text{where } j(\lambda', T) = \sum_{n=0}^{\infty} \frac{e^{-\lambda' T}(\lambda' T)^n}{n!}, \sigma_n^2 = \sigma^2 + \frac{n\delta^2}{T},$$

$$r_n = r - \lambda\hat{k} + \frac{nln(1+\hat{k})}{T}, \lambda' = \lambda(1+\hat{k}) \text{ and } q = r.$$

Both components can be seen here. $c_n(F_T, X, T, r_n, q, \sigma_n)$ (denoted as $c_n(\cdot)$ henceforth) represents the continuous drift whereby the asset price follows a GBM process with constant drift and volatility. $j(\lambda', T)$ (denoted as $j(\cdot)$ henceforth) describes the jump process, where λ is the jump intensity, $\hat{k}$ is the average jump process amplitude, δ is the variance of the jump process amplitude, and σ^2 is the variance of the diffusion process. A normal distribution is assumed for the jump size. The Merton model results in fatter tails than that of the Black-Scholes formula, a distribution that is more closely aligned to empirical asset prices observed. This is particularly evident over turbulent periods, such as those analysed in this study.

The effect of different parameter levels can be examined from a theoretical perspective. Parsing the jump and diffusion components and analysing these separately aids our understanding. Firstly, look at the situation when λ is equal to zero. This leads to a flat Black-Scholes implied volatility curve, as the expectation is that there will be no jump occurrences within the next year. For this reason, the jump component has no impact. Given that λ cannot be negative, the only other scenario is that λ is greater than zero. In this case, if $\hat{k}$ and δ are both zero, then a flat Black-Scholes implied volatility curve is also produced. There is positive expectation of a jump occurring but the jump size has an

expected value of zero with no variation, so jumps do not impact prices.

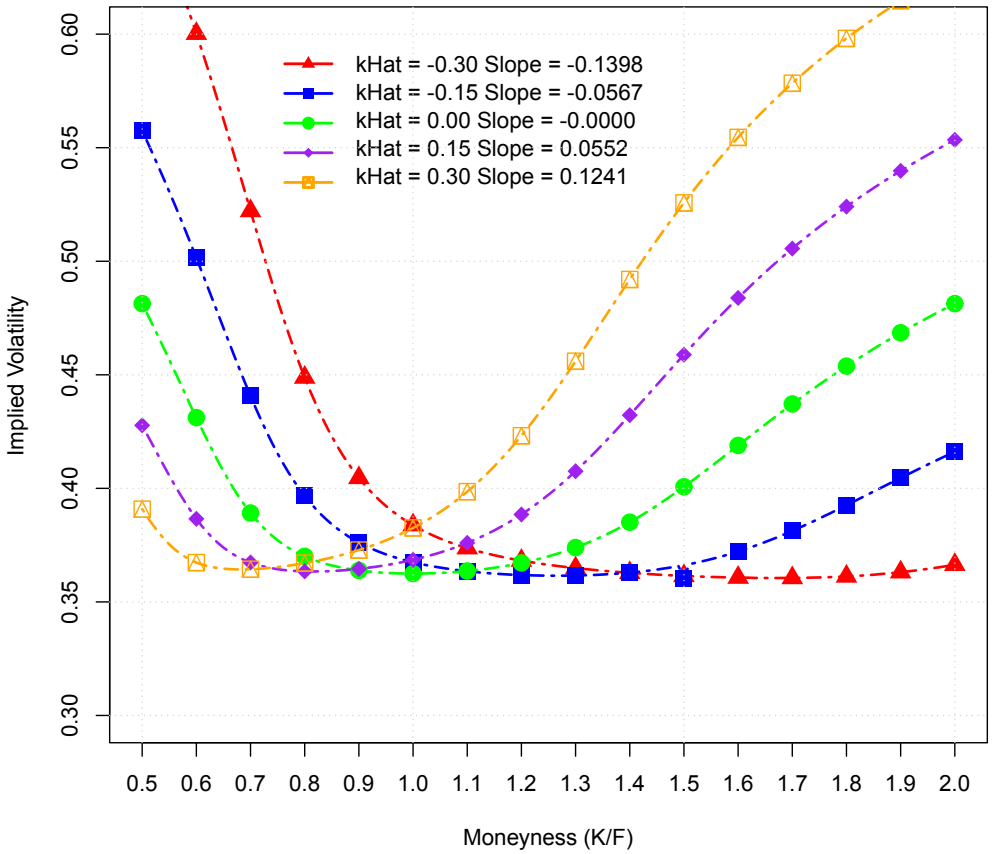
Some more interesting volatility curve shapes are seen when $\hat{k}$ is greater than zero. As the absolute value of $\hat{k}$ increases, the option value increases. This is due to a larger average jump size. The larger the specified $\hat{k}$, in either direction, the more jump occurrences affect the option price. This additional volatility results in a higher price as the probability of the option being exercised increases. As the $\hat{k}$ value increases, the higher orders of the summation term n have an increasing influence on $j(\cdot)$. The same dynamic is also true for $c_n(\cdot)$, as values associated with higher orders of n become larger as $\hat{k}$ increases. Given that the option value is a product of these terms, they interact to increase the option value. δ is the variance of the jump process amplitude and is incorporated in $c_n(\cdot)$, as an adjustment to the standard drift volatility input of the Black-Scholes model. There is a positive relationship between volatility and option price in the Black-Scholes formula. Therefore, an increase in δ increases the option value.

Other parameter inputs that affect the option price are the strike and futures prices. These are inputs of the Black-Scholes formula and do not influence the Poisson jump process. As a result, moneyness levels only impact $c_n(\cdot)$. The lower the moneyness (i.e., more ITM), the higher the option premium, as it is more likely to be profitable at the exercise date. An interesting dynamic is that higher orders of n impact the various moneyness levels of options differently. Higher order values of the summation term n have a greater impact on the pricing of OTM options as opposed to ITM options. This drives the dynamic that a move from negative to positive $\hat{k}$ results in $c_n(\cdot)$ increasing proportionately less for ITM values than for OTM values, whereas $\hat{k}$ being negative results in $c_n(\cdot)$ increasing proportionately less for OTM values than for ITM values. The effect of this is a skew when analysing the implied volatility curve associated with such prices. It results in a positive skew being observed for positive $\hat{k}$ values and a negative skew for negative $\hat{k}$ values.

Figure 3.2 graphically conveys the stylised facts for the $\hat{k}$ parameter in the Merton model. The red and blue lines on the graph represent the volatility curves when negative $\hat{k}$ parameters are specified, showing a negative skew in the volatility smile. There is a downward slope between ITM and ATM with relatively small increases in implied volatility as progression is made towards higher moneyness levels. A $\hat{k}$ value of zero leads to an almost symmetrical volatility curve as observed from the green curve with the Black-Scholes diffusion component driving the dynamic. Positive values for $\hat{k}$, however, the orange and purple lines on the graph, lead to a change in dynamic with a positive skew being observed, as the modelled expectation is that prices will rise.

To summarise, as downward jump fear increases, the ATM volatility smile slope becomes more negative and as upward jump fear intensifies, the ATM slope becomes more positive. This is exhibited through analysing both the empirical data graphs and a theoretical examination of the Merton call option pricing formula for differing levels. Indeed, the slope of the implied volatility curve is proportional to the jump size and direction. In line with the proxy employed for equity markets by Yan (2011) the relationship can be used to specify the average jump amplitude level in the Merton framework.

Figure 3.2: Implied volatility curve derived from various Merton model kHat levels



The Merton jump diffusion model is parameterised by δ=0.2, λ=0.3 and σ=0.35 and various $\hat{k}$ levels. The ATM slope is given in the legend.

Cont and Tankov (2004) and Giot et al. (2010) show that simultaneously optimising all the Merton model parameter can lead to a calibration problem that is ill-posed, even for simple jump diffusion models. In other words, there is a large range of parameter values that match observed market prices to within a reasonable error tolerance and calibration errors can have serious hedging consequences. Our approach is to select reasonable values of λ and δ and then extract the implied $\hat{k}$. $\hat{k}$ is of particular importance as it indicates both the market expectation of jump size and jump direction. In the selection of reasonable values for λ and δ, the heuristic approach of Murphy and Ronn (2015) is followed. They employ a Ball and Torous (1983) analysis of the same underlying futures data set and calculate the average number of jumps above three, four, and five standard deviations to set λ to 0.3. Again using mean values in line with Ball and Torous (1983), δ is set at 0.1. The focus of this paper is on the market implied jump size and direction from a relative historical setting, and fixing λ and δ to appropriate values in this regard, does not significantly impinge on the results.

3.2.4 Delta hedging application

A key innovation of this paper is the application of FDA techniques to portfolio management, in particular a delta hedging example. The hypothesis is that the use of FDA analysis to obtain parameter levels for the Merton model leads to superior hedging performance. The cumulative next day profit and loss is used to measure the cost of implementing the hedge with performance compared to that of the standard Black-Scholes delta strategy in line with Murphy and Ronn (2015). Moschini and Myers (2002) outline how important it is to hedge in commodity markets. First, a portfolio is constructed consisting of long one call option and short delta underlying future contracts. Utilising the following day's closing option and future prices, the profit and loss of the portfolio is calculated. The hedging error is also recorded, which is the absolute value of this daily profit and loss figure. It measures how closely the hedge tracks the portfolio, regardless of whether the hedge represents a profit or loss. This procedure is repeated for each day in the sample with both the cumulative profit and loss, as well as the cumulative hedging error being calculated.

Two different delta hedging ratios will be calculated and compared:

1. The widely implemented delta as calculated from the d_1 component of the Black-Scholes formula:

$$d_1 = \frac{ln(\frac{S}{K}) + (r + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}}.$$

   Delta is equal to $N(d_1)$, where $N(\centerdot)$ is the cumulative distribution function of the standard normal distribution.

2. The second delta hedging ratio is based on the Merton delta algorithm derived by

Murphy and Ronn (2015):

$$\triangle = \frac{\frac{\delta c}{\delta S} \mid_{\lambda=0} \sigma^2 S^2 \triangle t + \frac{\delta c}{\delta S} \mid_{\sigma=0} \lambda \delta^2}{\sigma^2 S^2 \triangle t + \lambda \delta^2},$$

where $\frac{\delta c}{\delta S} \mid_{\lambda=0}$and $\frac{\delta c}{\delta S} \mid_{\sigma=0}$ are taken from the Merton model.

3.3 Data set

The data have been downloaded on a daily basis from the CME Group's FTP site. The data set spans from 2 April 2007 to 31 January 2013 and includes traded WTI call oil option price, underlying future price, maturity length, strike, implied volatility, and the associated discount rate on each trading day. All option quotes are subject to the following screening criteria:[11]

- The option must have a minimum dollar value of \$0.05. Options with such a low market value might display differing characteristics due to very little active trading. Illiquidity can distort the data.
- $0.5 \leq$ Strike Price/Futures Price ≤ 2.0, i.e., it must trade between these two moneyness bounds. This is to ensure sufficient liquidity as there are fewer markets participants at extreme moneyness levels.
- Options have a maturity date of between 1 and 2 months (WTI CL02). This is the most frequently traded maturity contract available.

3.4 Empirical results

3.4.1 Impact of economic factors on implied volatility

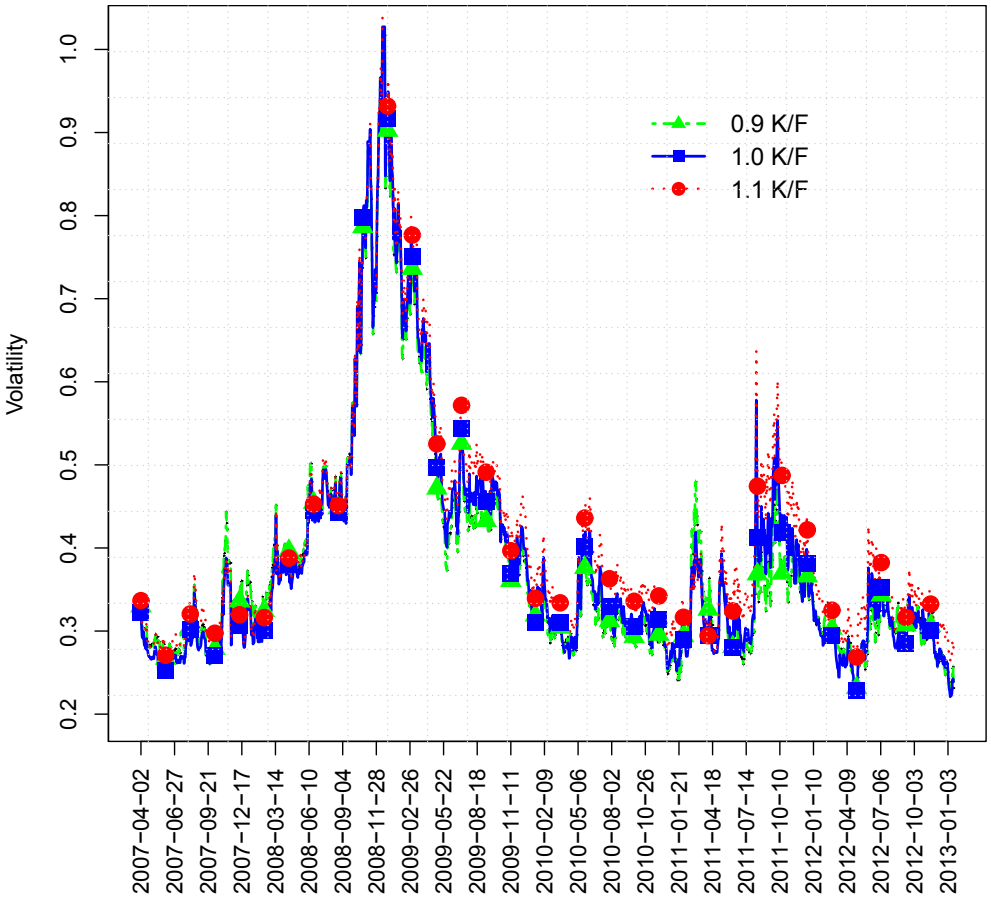
Fourth order B-spline basis functions are applied to reduce the dimensionality of the continuous process underlying the daily implied volatility curves.[12] Each curve is defined over the range of moneyness values in the data set, [0.72, 1.24]. To construct the functional data objects, a smoothing parameter value, λ^*, of 0.001 is selected. The plot of the values of the GCV criterion presented in Figure 1 of the Appendix is used as an indicator.[13] Functional data objects of implied volatility curves for each day over the five-year period

[11]The historical data has kindly been made available to the authors by University of Texas at Austin for which they are extremely grateful.

[12]The fda package in R is used for the analysis.

[13]GCVs are calculated for a number of other periods in the sample, with no material differences observed.

Figure 3.3: Crude oil implied volatility over time [2007-2013]

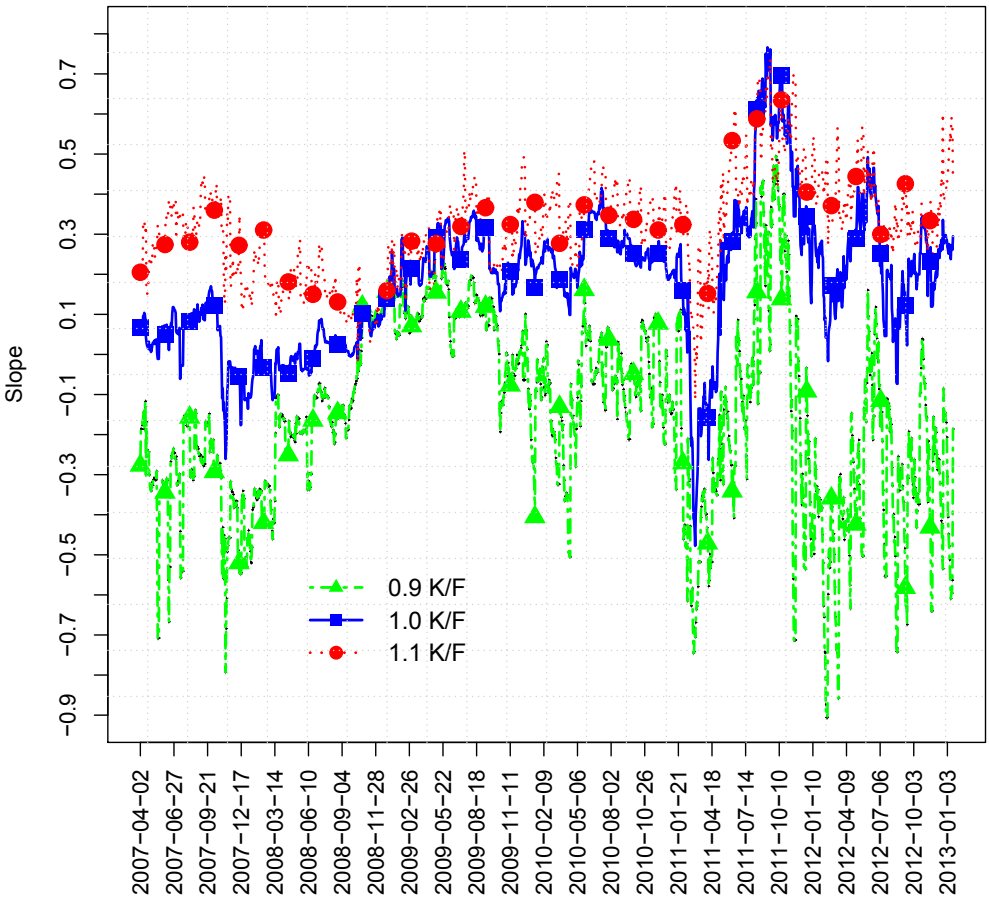


FDA obtained implied volatility curve values at three points—0.9 moneyness, 1.0 moneyness, and 1.1 moneyness—plotted daily between April 2007 and January 2013.

are shown in Figure 3.3. This representation allows us to obtain slope and curvature values directly. The slope and curvature values are evaluated at three points along the moneyness curve, namely ATM, 10% ITM, and 10% OTM. Figure 3.4 presents the dynamics of the slope values, with Figure 3.5 showing the curvature levels through time. These graphs lead an analysis of the implied volatility curve shape in terms of the supply and demand side weaknesses observed during the 2007-2013 period. The slope values for each moneyness level in Figure 3.4 are first analysed in order to obtain an understanding of the shape of the implied volatility curve.

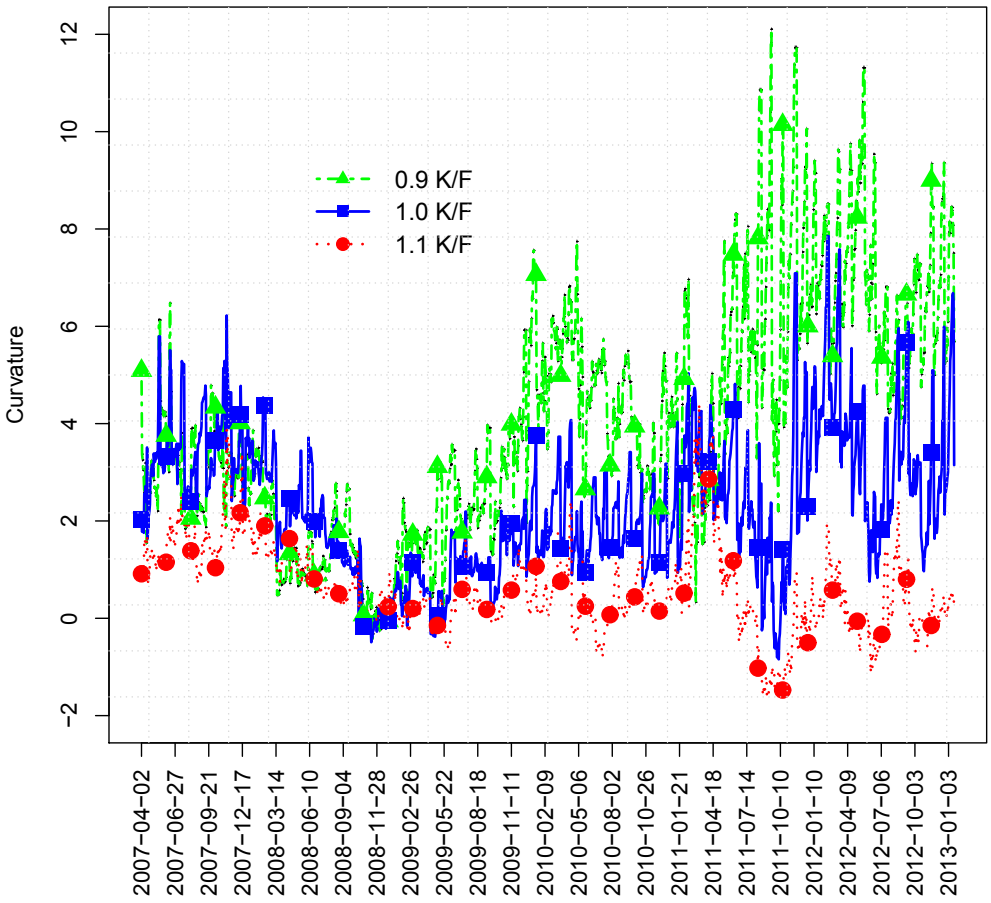
During the benign pre-financial crisis period of 2007, the ATM slope levels fluctuate around zero, signifying that the turning point of the smile is located in this range. Concurrent with this, the ITM slope is negative while the OTM slope is positive, in agreement with a typical U-shaped volatility smile. The fear in crude oil markets during benign periods is that of a price spike, so the prevailing expectation is that the price will be higher by the expiration date. This should lead to a positive skew being demonstrated through larger

Figure 3.4: Crude oil implied volatility slope 2007-2013



FDA obtained implied volatility curve values at three points—0.9 moneyness, 1.0 moneyness, and 1.1 moneyness—plotted daily between April 2007 and January 2013.

Figure 3.5: Crude oil implied volatility curvature 2007-2013



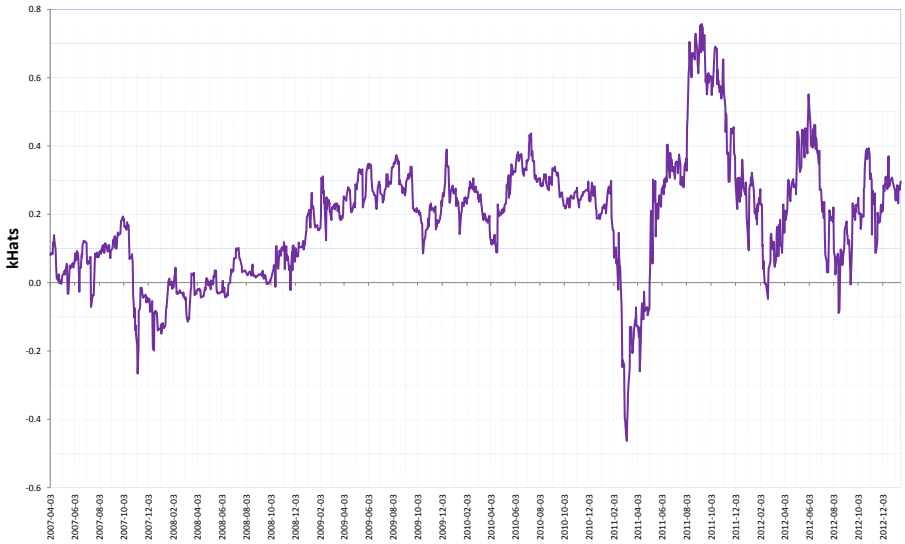
FDA obtained implied volatility curve values at three points—0.9 moneyness, 1.0 moneyness, and 1.1 moneyness—plotted daily between April 2007 and January 2013.

absolute OTM slope levels than absolute ITM slope levels. However, this is not evident in the lead up to the recent global financial crisis and could perhaps be interpreted as an early indicator of demand side weakness, where the expectation is that of a price drop. The recent global financial crisis is a period of extreme volatility levels as seen in Figure 3.3. It brings about a large drop in demand, resulting in a flatter volatility smile. This is demonstrated through the three slope levels in Figure 3.4 converging towards zero. During this period, the elevated level of volatility across the moneyness spectrum dominates any nuances concerning the skew. This is in contrast to Chuang et al. (2013) who find that implied volatility curves in equity markets are less skewed when volatility levels are lower.

As the extreme volatility levels of the financial crisis slowly dissipate, a small positive ATM slope level, a negative ITM, and a positive OTM slope are observed in Figure 3.4. The negative ITM and positive OTM slope in this 2009-2010 period suggest convergence to a more conventional implied volatility curve shape. Furthermore, the absolute value of the OTM slope is higher than that of the ITM slope, indicating a return of the positive skew that is typical of commodity options. The first full year of the Arab Spring in 2011 results in a dramatic change, however, as volatility levels, shown in Figure 3.3, increase once again. Tunisia is the source of the revolution outbreak beginning in late December 2010 but never experiences a significant disruption in crude oil production. The prevailing fear in the early stage, circa January 2011-March 2011, is that of demand side weakness due to escalating security anxiety. The three curves in Figure 3.4 exhibit this through each slope level shifting downwards in tandem. The ITM negative slope steepens and the OTM positive slope flattens. This dynamic more closely represents the negative skew observed in equity markets where the predominant fear is that of a price crash. March 2011 sees Libya experience a full-scale war, resulting in long-lasting logistical disruptions to crude oil exports and potential for further unrest. The remainder of 2011 is dominated by a contagion effect across the region, with related protests erupting in other key crude oil producers in the region, such as Bahrain, Iraq, Saudi Arabia, Kuwait and Syria. The shape of the volatility smile acts accordingly with positive ATM, ITM, and OTM slope values. Lower levels of volatility are observed in the 2012-2013 period, as demonstrated in Figure 3.3. This corresponds with Figure 3.4 exhibiting a return to a more typical commodity smile slope whereby the ITM slope is negative, the OTM slope is positive, and the ATM slope registers a small positive level too. The fear of a price spike is still evident here, however it has quelled somewhat from the extreme positive skew observed during the height of the Arab Spring in 2011.

The focus of the analysis now shifts to the curvature plot in Figure 3.5. Analysing curvature levels for the three moneyness ranges collectively, highlights the underlying economic dynamics. The scale of the curvature levels increase over time due to greater market turbulence, with large absolute values, as high as +12 for ITM, and as low as -2 for OTM, registered in the latter years of the sample. The curvature values are predominantly positive, which, indicate a function that is predominantly concave down (convex). This is consistent with a general increase in the slope magnitude across the moneyness spectrum, as shown in Figure 3.3. A strangle is an option strategy whereby the holder benefits from a large movement away from the current underlying price, in either direction. The implied volatility curvature is a measure of the strangle premium, as it indicates how much the two

Figure 3.6: $\hat{k}$ proxy used over time



The Merton (1976) jump diffusion model $\hat{k}$ input parameter values obtained using the FDA obtained ATM slope.

OTM strangle volatilities are above the ATM volatility (Beber et al. 2010). The contrast between the lower underlying volatility levels and the increasing implied volatility curvature can be interpreted as a strangle market, with participants placing a higher probability on a large move in either direction and demand driving the strangle premium upwards.

When coupled with differing slope levels, the primarily positive curvature values lead to quite different dynamics during the benign and crisis periods in the sample. Relatively low curvature levels are seen across the moneyness spectrum during both crisis periods, be it the financial crisis or the Arab Spring. This is in comparison to benign times in the lead up and aftermath of such shocks, whereby the implied volatility slope changes become significantly less sharp through the ITM range and show sharper slope changes through the ATM and OTM points.

3.4.2 Delta hedging performance

We now compare the performance of delta hedging approaches utilising both the Black-Scholes and our FDA-optimised Merton model. The Black-Scholes and optimised Merton deltas are calculated and used to get the next day profit and loss for a strategy that is long one ATM strike call option and short delta times the underlying future contract. Given the linear relationship between the $\hat{k}$ parameter of the Merton model and the implied volatility curve's ATM slope established in the previous section and by Yan (2011), the ATM slope is used as a proxy for average jump amplitude.

Figure 3.6 shows the average jump amplitude levels obtained utilising our proposed proxy. Over the entire sample, the majority of $\hat{k}$ values observed are positive. This can be viewed as a signal that even during turbulent times the predominant fear in crude oil

markets is that of a price spike. The $\hat{k}$ magnitude increases with the passing of time, fluctuating in the -0.2 to +0.2 range up to the beginning of 2009, with extreme values between -0.5 and +0.8 exhibited in 2011. Significant changes in $\hat{k}$ levels are an indicator of different underlying economic forces. Beginning at the start of the sample period, April 2007-November 2007 shows small positive $\hat{k}$. November 2007-January 2009 corresponds to a predominantly negative $\hat{k}$ as a result of the demand side weakness brought about from the global financial crisis and its lead-in period. An extended period, January 2009-January 2011, of positive and relatively stable $\hat{k}$ is seen as indicating a more benign time within crude oil markets. The onset of the Arab Spring in late 2010 however, shows a sharp drop in the level of $\hat{k}$, with values as low as -0.5 in February 2011. The initial prevailing fear in crude oil markets being that of demand side weakness, brought about due to escalating security anxiety in the Middle East. As the movement spreads, disruptions to supply are deemed to be more probable. This is represented by increasing $\hat{k}$ values in Figure 3.6. March 2011 saw Libya experience a full-scale war, resulting in long-lasting logistical disruptions to crude oil exportation and the potential for further unrest, leading to high positive $\hat{k}$ figures, such as the +0.8 level in late 2011. The contagion effect across the region accelerated the shift from negative to positive $\hat{k}$, with related protests erupting in other key oil producers in the Middle East, such as Bahrain, Iraq, Saudi Arabia, Kuwait, and Syria. The 2012-2013 period following the uncertainty of the Arab Spring corresponds to high $\hat{k}$ as the market corrects prices to pre-supply shock levels coupled with the restoration of economic growth. These period breakdowns are used to provide a more granular view of the sample's hedging performance.

The results of the analysis are summarised in Table 3.1. When implementing a hedge, a key consideration for the portfolio manager is how much the strategy costs, so the primary focus here is the profit and loss figure, which is calculated daily and aggregated over the period. The Merton model delta hedging strategy, optimised with the ATM slope derived $\hat{k}$ parameter, outperforms the Black-Scholes strategy over the entire period, with a mean daily loss of only 0.0069 compared to 0.0075. This gives an indication that our optimised Merton strategy is preferable to the commonly implemented Black-Scholes based strategy as it equates to an 8% reduction in cost. The daily variation of the profit and loss for both strategies is very similar at 1.0373 for our optimised Merton strategy versus 1.0365 for the Black-Scholes model. When looking at the hedging error over the entire sample period, there is very little difference between both strategies also with our optimised Merton model showing a hedging error of 0.7678 with 0.7673 seen for the Black-Scholes model. Therefore over the entire sample it can be concluded that our Merton strategy provides superior performance. With a view to identifying the implied volatility shape under which the Merton model outperforms the Black-Scholes, the sample is broken down into periods of different underlying market activity, split by $\hat{k}$ variation; April 2007-November 2007, November 2007-January 2009, January 2009-January 2011, February 2011-May 2011, and finally May 2011-December 2011.[14] First, the pre-financial crisis months of April 2007-

[14]The December 2011 to January 2013 period is not listed as it constitutes a period of multiple positive

Table 3.1: Delta hedging results

Period	BS Delta PnL	Merton Delta PnL	BS Hedge Error	Merton Hedge Error	$\hat{k}$
Apr 2007-Jan 2013					
Mean	-0.0075	-0.0069	0.7673	0.7678	0.1828
SD	1.0365	1.0373	0.6966	0.6973	0.1766
Max	5.1935	5.1654	6.6877	6.7011	0.7570
Min	-6.6877	-6.7011	0.0004	0.0003	-0.4634
Apr 2007-Nov 2007					
Mean	-0.0610	-0.0557	0.4791	0.4788	0.0771
SD	0.5991	0.6006	0.3626	0.3647	0.0577
Nov 2007-Jan 2009					
Mean	0.0952	0.0953	1.0968	1.0970	-0.0011
SD	1.4243	1.4241	0.9115	0.9108	0.0700
Jan 2009-Jan 2011					
Mean	-0.0534	-0.0523	0.6550	0.6549	0.2548
SD	0.8443	0.8438	0.5346	0.5339	0.0607
Jan 2011-May 2011					
Mean	-0.1860	-0.1875	0.7166	0.7184	-0.1192
SD	0.9235	0.9257	0.6051	0.6069	0.1483
May 2011-Dec 2011					
Mean	0.0468	0.0467	0.9613	0.9624	0.4307
SD	1.2697	1.2708	0.8273	0.8278	0.1795

The recorded performance of implementing both a Merton (1976) jump diffusion model derived delta hedging strategy, and a standard Black-Scholes (1973) delta hedging strategy, between January 2007 and January 2013. Mean is the average daily value. SD is the standard deviation of this value. PnL is an abbreviation for Profit and Loss and Hedge Error is the absolute value of the daily PnL.

November 2007 were a relatively benign period in the market where the risk is assessed as being that of a small upward spike in prices. This is shown in Figures 3.3 and **3.4**. Table 3.1 exhibits a daily mean profit and loss of -0.0557 versus -0.0610 and an average hedging error of 0.4791 versus 0.4788 for our optimised Merton strategy and Black-Scholes models respectively. This indicates both lower cost and lower absolute hedging error in this period for our optimised Merton strategy. During the demand side weakness of the global financial crisis (2007-2009), it can be seen in Table 3.1 that a profit is made through the implementation of either strategy. These profits are almost identical at 0.0952 and 0.0953 for the Black-Scholes and our optimised Merton model, respectively. This dynamic is seen throughout the sample, during periods in which there is a cost to implementing the Black-Scholes hedging strategy, the Merton strategy outperforms. These periods correspond to positive implied volatility slopes/$\hat{k}$ values. However, during periods where a profit is made implementing the Black-Scholes hedge the Merton strategy broadly matches its performance.

Each investor has individual hedging performance requirements, which of course take precedence in the decision of which strategy to implement.[15] That said, our FDA optimised Merton delta hedging strategy outperforms the Black-Scholes delta hedging strategy by 8% in terms of implementation cost, over the entire sample. Breaking the sample down into periods split by predominantly positive and predominantly negative implied volatility slopes, we see that the Merton strategy outperforms the Black-Scholes when $\hat{k}$ values are positive and broadly matches its performance in periods of negative $\hat{k}$ values.

3.5 Conclusion

The entire set of price dynamics within crude oil markets cannot be fully represented by traditional multivariate analysis. We combine the application of FDA techniques, jump components from the Merton model, and an analysis of underlying economic pressures to better understand market implied volatility and jump dynamics in the 2007-2013 sample period. The analysis should be of interest to both academics and market participants seeking to understand prevailing implied volatility, jump expectations, and crucially, jump direction. The study contributes through three major findings.

Firstly, we demonstrate that the implied volatility smile exhibits a positive skew in periods of supply side weakness and a negative skew in periods of demand side weakness. We clearly ex-post demonstrate the link between implied volatility shape and contemporary socio-economic events, especially during the turbulent years we examine. In our sample, the systematic analysis of implied volatility also highlights periods of economic weakness in advance of their occurrence. Secondly, we provide both theoretical and empirical evidence establishing a relationship between implied volatility shape information and jump amplitude for crude oil options, in a similar manner to what Yan (2011) has shown to be the case for stock returns. We achieve this by combining the constructed functional data objects

and multiple negative $\hat{k}$ values.

[15] Hammoudeh et al. (2013) emphasise the increased importance of risk management in volatile environments.

with key attributes of the Merton model to derive implied values for the average jump amplitude. Finally, we demonstrate how information contained in the implied volatility smile slope can be exploited to improve portfolio hedging techniques. Our FDA optimised Merton delta hedging strategy outperforms the Black-Scholes delta hedging benchmark by 8% in terms of implementation cost over the entire sample. Breaking the sample down into periods split by predominantly positive and predominantly negative implied volatility slopes, we see that the Merton strategy outperforms the Black-Scholes when $\hat{k}$ values are positive and broadly matches its performance in periods of negative $\hat{k}$ values.

Chapter 4

Forecasting implied volatility in foreign exchange markets: A robust functional linear model approach

4.1 Introduction

The Black-Scholes (1973) model assumes that volatility is constant. This assumption, if true, should result in a flat implied volatility curve; the market's expectation of average price volatility for the underlying asset to an option contract between now and its expiry date. Of course in practice, observed implied volatility differs across option contracts, dependent on both moneyness and expiry date. As well as being a transformation of the option price, and a key parameter in many asset pricing formulae, implied volatility is also of interest due to its informational content (see Corrado and Miller 2006, Taylor et al. 2010, Muzzioli 2010, and Garvey and Gallagher 2012). Yu et al. (2010) demonstrate this by finding superior results using implied volatility to predict future return volatility of stock index options, when compared to traditional benchmark models in over-the-counter (OTC) and exchange markets. One such OTC market is that of foreign exchange (FX) options. FX is the largest asset class in the world with the Bank for International Settlements reporting that trading levels in FX markets averaged \$5.3 trillion per day.[1] Many stakeholders are exposed to FX risk including banks, speculators, traders, multinational firms, importers, and exporters. Modelling foreign currency cash flows, investment decisions, and hedging strategies, are all greatly dependent on expectations of future FX movements. Our study adds to the existing literature through the novel proposal of a functional data analysis-based forecasting model to predict the evolution of the implied volatility function. The aim is to determine and forecast the function that characterises the implied volatility relationship among option contracts. We not only contribute from an academic perspective, where insights into the dynamics of implied volatility aid our understanding of option

[1] http://www.bis.org/publ/rpfx13fx.pdf

markets, but also from a market practitioner perspective, due to the study's potential hedging and speculation implications.

Compared to previous studies forecasting the volatility of returns, there is a relative paucity of literature predicting the evolution of implied volatility. Examples include Goncalves and Guidolin (2006), Konstantinidi et al. (2008), Chalamandaris and Tsekrekos (2010), Dunis et al. (2013), Bernales and Guidolin (2014), and Chalamandaris and Tsekrekos (2014). Konstantinidi et al. (2008), for instance, use a number of economic indicators to construct a forecasting model that finds statistically significant predictable patterns in the evolution of European and U.S. implied volatility indices. Dunis et al. (2013) apply the same economic model to predict the evolution of implied volatility in the EUR-USD exchange rate, a currency pair which we also study. They find that implied volatility is only predictable at short time horizons of up to 5 hours ahead. Chalamandaris and Tsekrekos (2014) study Euro OTC FX options, and find that none of their proposed implied volatility models consistently outperform the autoregressive benchmark in short horizon forecasts, of less than 5 days ahead. They also conclude that structured parametric implied volatility forecasting models lead to superior out-of-sample results, a conclusion that we seek to disprove through the use of a flexible functional data model. Such a functional approach aims to uncover the process underlying a data set and incorporates shape into its forecast. It offers infinite dimensional space representation which exposes additional dynamics missed by traditional multivariate techniques. Furthermore, functional data analysis boasts the advantages of being computationally efficient and of allowing functions to be evaluated on an arbitrarily fine grid. These and other advantages of FDA are outlined in Ramsay and Silverman (2005).

A number of applications in the bio-mechanical literature have incorporated functional data analysis. However, it has only recently been exploited for financial analysis. Muller et al. (2011) study high frequency S&P 500 Index levels, and propose a functional volatility process to model volatility trajectories. Their model shows patterns in volatility and by combining it with prediction techniques and functional regression, it can be used to predict future volatility. Benko et al. (2009) focus on implied volatility, by presenting a new two-sample common factor FPCA technique and applying it to analyse similarities in stochastic behaviours between implied volatility curves of one- and three- month option contracts on the German-Swiss exchange (EUREX). They highlight the strength of using functional data analysis techniques to characterise the implied volatility function, an approach which we also adopt.

Our study is distinct, in that we use a functional linear model to obtain superior out-of-sample forecasts. Both the scalar response/functional explanatory and functional explanatory/functional response linear models of Ramsay and Silverman (2005) are utilised for the analysis. We show that these models outperform traditionally proposed AR, GARCH, and ARFIMA benchmarks with the results being statistically significant in out-of-sample testing. We contribute by incorporating the use of a contributory data vendor. This mitigates the idiosyncratic risk, as highlighted by Chalamandaris and Tsekrekos (2014), associated

with obtaining quotes from a single market participant.

We contribute further by incorporating controls for the multiple comparisons problem in our forecasting framework. This robust testing framework adjusts for the likelihood that *seemingly* significant ouperformance can be due to mere chance alone. As the number of simultaneous tests conducted increases, so too does the likelihood of such false discoveries. This issue is known as the multiple comparisons problem and must be controlled for when studying forecasting performance. To solve this issue we implement the operative balanced stepdown procedure of Romano and Wolf (2010), the first time it has been applied in the volatility forecasting literature.[2] The balanced stepdown procedure offers a more generalised and flexible approach to controlling for the multiple comparisons problem than previous frameworks proposed. The methodologies used in previous implied volatility forecasting studies raise concerns around the validity of the inferences drawn, insofar as many lack multiple comparisons controls. We demonstrate intertemporal dependency across the moneyness range, as well as implied volatility predictability in the highly liquid EUR-USD pair that we study. The results are of interest to both academics, given potential market efficiency implications, and market practitioners, who may seek to exploit the uncovered patterns. The remainder of the paper is organised as follows. Section 5.3 provides a background to the functional data analysis methodology and the forecast evaluation procedure. Section 5.3.4 details the multiple comparisons problem and the Romano and Wolf (2010) operative balanced stepdown procedure. Section 4.4 introduces the EUR-USD FX options data set. Section 4.5 presents and discusses the empirical results, with Section 5.5 concluding the paper.

Summary of contributions

- Can functional linear model techniques be used to characterise and forecast implied volatility in foreign exchange markets?
- How does the performance of the functional data analysis approach compare to traditionally employed benchmark models?
- Are the findings robust across various moneyness segments, contract maturities and out-of-sample window lengths?

4.2 Methodology

4.2.1 Functional representation

Functional data analysis (FDA) provides a functional representation of the process underlying a data set. The process is defined over a continuum of values, where the

[2]The application of the balanced stepdown procedure of Romano and Wolf (2010) is in line with Cummins and Bucca (2012) and Kearney et al. (2014), who adopt the framework in the identification of profitable statistical arbitrage opportunities and exchange traded fund outperformance, respectively.

continuum is generally represented in terms of time or space. In this paper the functions are defined in the moneyness domain, as we characterise and forecast the evolution of the implied volatility process. As will be outlined in this section, the FDA methodology has many advantages; it accurately captures implied volatility dynamics (Benko et al. 2009), there is no assumed parametric structure, it is computationally efficient, and it results in a process that can be evaluated on an arbitrarily fine grid. These and other advantages of FDA are outlined in Ramsay and Silverman (2005).

Using daily observed option implied volatility data, $x(m)$, we uncover a functional data object, or more simply, the function, denoted $\tilde{x}(m)$, that determines the daily implied volatility curve dynamics. The domain, m, is that of the moneyness level (in terms of delta). When constructing a functional data object, a vector of n bases, denoted $\phi_1, ..., \phi_n$, must first be specified. The decision of which basis system to specify is driven by the underlying data's known characteristics. For instance, when modelling periodic data, a Fourier basis expansion, comprised of successive sine/cosine terms, is most commonly applied. However, an implied volatility process does not exhibit strong cyclic variation, so we choose B-splines for the basis function system. B-spline representation offers a number of strengths, as outlined in de Boor (2001). Computations with B-splines are extremely efficient as at any one point along the curve they simplify to a polynomial that can be easily evaluated. Adjusting the order of the spline allows for the estimation of derivatives of any degree. In this paper, a fourth order basis, or cubic polynomial is specified. Specifying a cubic polynomial provides a good balance as it retains the function's continuous property up to the second derivative. B-splines are essentially a number of polynomials joined together smoothly at fixed points called knots. The number and positioning of the knots are derived from knowledge of the complexity of the underlying process over particular ranges. We place knots at the discrete quoted option moneyness levels available from the data set, with polynomials describing the moneyness interval between the knots. This results in q knots: $m_k : m_1 \leq ... \leq m_q$, with the range of the various sub-intervals, $[m_k, m_{k+1}]$, being defined through the placement of these knots. Within each sub-interval, the spline is simply a polynomial of order n. The order is calculated as:

$$order = 1 + degree\, of\, the\, polynomial.$$

Functional structures are approximated as a weighted linear combination of these bases:

$$\tilde{x}(m) \equiv c_1\phi_1(m) + c_2\phi_2(m) + ... + c_n\phi_n(m),$$

where $c_1, ..., c_n$ represent the parameters of the expansion's coefficients. As in Ramsay and Silvermann (2005), the coefficients c_j can be chosen by minimising:

$$SSE(c_1, ..., c_n) \equiv \sum_{k=1}^{q} [x(m_k) - \tilde{x}(m_k)]^2 = \sum_{k=1}^{q} \left[x(m_k) - \sum_{j=1}^{n} c_j\phi_j(m_k) \right]^2 \qquad (1)$$

where SSE stands for the sum of squared errors and q represents the number of implied volatility observations.

4.2.1.1 Smoothing parameter

To avoid over-fitting the data, a smoothing penalty is applied in calculating the basis coefficients of the implied volatility process. The smoothing penalty helps to remove noise from the data. Noise may be present due to liquidity issues, misquotes or other data irregularities masking the true function. Without smoothing, noise in the relatively small number of discrete values available can distort the results for that range. This can be particularly evident at extreme moneyness levels. In line with Ramsay and Silvermann (2005) and Liu et al. (2012), a limitation is placed on the variation of the curvature. The total curvature of the process is found by integrating its squared second derivative:

$$R(\tilde{x}) \equiv \int \left(\frac{d^2}{dm^2} \tilde{x}(m) \right)^2 dm.$$

This is also called the roughness of the function.

In an extension of (1), the coefficients characterising the smoothed curve are found using the penalised sum of squared errors:

$$PENSSE(c_1, ..., c_n) \equiv \sum_{k=1}^{q} \left[x(m_k) - \sum_{j=1}^{n} c_j \phi_j(m_k) \right]^2 + \lambda R(\tilde{x}).$$

As λ increases, more weight is placed on the roughness penalty, possibly missing some of the process' dynamics. As λ decreases, less weight is placed on the roughness penalty and only data fitting matters in uncovering the function. In order to balance the competing goals of retaining features and removing noise from the data, an optimal smoothing level, λ, must be selected. Using generalised cross validation developed in Craven and Wahba (1979), and adopted by Ramsay et al. (2009), and Liu et al. (2012), we select $\lambda = 10^{-3}$.

4.2.2 Functional linear model

A functional linear model is utilised to predict the evolution of the implied volatility process. Classical linear models seek to describe the dependency between a response variable and a specified set of predictors. In classical regression, scalar values are used for both the explanatory and response variables. However, in functional linear regression at least one of the observed explanatory variables are curves. This means that functional analogs of classical linear regression coefficients must be constructed. The procedure varies according to the model structure. Given that the explanatory variable adopted in our study is the implied volatility function we employ the use of:

1. Scalar response/functional explanatory which takes the form $y = \alpha + \int \beta(m)\tilde{x}(m)dm + \varepsilon$ (Hovarth and Kokoszka 2012) ("scalar response model", henceforth)

2. Functional response/functional explanatory which takes the form $y(m) = \alpha(m) + \int \beta(m,s)\tilde{x}(s)ds + \varepsilon$ (Hovarth and Kokoszka 2012) ("fully functional model", henceforth)

The forthcoming sections discuss these two models in detail.

4.2.2.1 Scalar response model

We utilise the scalar response framework to find the dependency between the current day, t, implied volatility function, $\tilde{x}_t(m)$, and the one-day ahead, $t+1$, implied volatility scalar response for a particular contract, $x_{t+1}(m_k)$:

$$x_{t+1}\left(m_k\right) = \alpha + \int_{\Omega_m} \beta(m)\tilde{x}_t(m)dm + \varepsilon_t,$$

where Ω_m is the defined moneyness range, and where $\hat{\beta}(m)$ is found by minimising:

$$\sum_{t=1}^{T-1}\left(x_{t+1}(m_k) - \alpha - \int_{\Omega_m} \beta(m)\tilde{x}_t(m)dm\right)^2. \qquad (4)$$

In classical linear regression, there must be fewer explanatory variables than observations. Using a functional explanatory variable, however, acts as an infinite-dimensional predictor of a finite set of responses. This means that an exact fit is always possible, leading to $\varepsilon = 0$. It also means that an infinite number of possible $\beta(m)$ coefficients will produce the same predictions. Dimension reduction through a basis expansion of $\beta(m)$, as in Section 4.2.1, is proposed by Ramsay and Silvermann (2005) to solve this underdetermination issue. The smaller the number of basis functions, the smoother the estimate function $\hat{\beta}(m)$. However, a low-dimensional basis may not be appropriate as it has the potential to omit important dependency dynamics. To allow for the use of a high-dimensional basis, $\hat{\beta}(m)$ can be smoothed to obtain an appropriate estimate for the continuum-varying coefficient $\beta(m)$. This is done by imposing a roughness penalty which minimises deviations from $\frac{d^2}{dm^2}\hat{\beta}(m) = 0$. After incorporating the penalty, a smoothed $\hat{\beta}(m)$ is found by minimising:

$$\sum_{t=1}^{T-1}\left(x_{t+1}(m_k) - \alpha - \int_{\Omega_m} \beta(m)\tilde{x}_t(m)dm\right)^2 + \lambda_\beta \int \left[\frac{d^2}{dm^2}\beta(m)\right]^2 dm,$$

where λ_β is the weighting attributed to the smoothing penalty. Given that 5 and 95 represent the lower and upper bound delta values in the data set, we can define our model as:

$$x_{t+1}(m_k) = \alpha + \int_5^{95} \beta(m)\tilde{x}_t(m)dm + \varepsilon_t.$$

4.2.2.2 Fully functional model

We utilise the fully functional model as an exploratory tool only to assess the dependency between the current day, t, implied volatility function, $\tilde{x}_t(m)$, and the one-day ahead, $t+1$, implied volatility function, $\tilde{x}_{t+1}(m)$. Given that both variables are expressed in terms of moneyness, we use the notation m and m' to distinguish between the moneyness domains of the current day, t, and the next day, $t+1$ implied volatility functions, respectively. We specify a fully functional model based on the historical linear framework proposed by Malfait and Ramsay (2003):

$$\tilde{x}_{t+1}(m^{'}) = \alpha(m') + \int_{\Omega_m} \beta(m, m')\tilde{x}_t(m)dm + \varepsilon_t(m') \qquad (5)$$

where Ω_m contains the domain range of m over which $\tilde{x}_t(m)$ is considered to influence $\tilde{x}_{t+1}(m')$. We predict $\tilde{x}_{t+1}(m')$ using the entire range of the $\tilde{x}_t(m)$ function, i.e., 5 to 95 delta.

In a similar view to the scalar response model, dimension reduction through a double basis expansion of $\beta(m, m')$, in terms of both m and m', is used to solve the underdetermination issue. The smaller the number of basis functions, the smoother the estimate function $\hat{\beta}(m, m')$. However, two low-dimensional bases may not be appropriate as they have the potential to omit important curve dynamics. To overcome this issue, Ramsay and Silverman (2005) apply an additional roughness penalty, to smooth in terms of both the range specified by m and m'. Weightings for the penalties are defined as λ_1 and λ_2, with the penalty being structured as follows:

$$\lambda_1 \int \left[\frac{\partial^2}{\partial m'^2}\beta(m, m')\right]^2 dmdm' + \lambda_2 \int \left[\frac{\partial^2}{\partial m^2}\beta(m, m')\right]^2 dmdm'.$$

Given that the specified explanatory and responses are both curves, the resultant $\hat{\beta}(m,m')$ value takes the form of a 3-dimensional surface object, which we present in Section 4.5.

In order to assess how well the functional models fit the data, functional versions of the widely employed R^2 statistic and F-Ratio are applied:

$$R^2(m) = 1 - \frac{\sum_t^{T-1}(x_{t+1}(m_k) - \hat{x}_{t+1}(m_k))^2}{\sum_t^{T-1}(x_{t+1}(m_k) - \bar{x}_{t+1}(m_k))^2}$$

where $x_{t+1}(m_k)$ is the observed response, $\bar{x}_{t+1}(m_k)$ is the mean of the observed response, and $\hat{x}_{t+1}(m_k)$ is the model's estimated response value.

$$F - Ratio = \frac{(\sum_t^{T-1}(x_{t+1}(m_k) - \bar{x}_{t+1}(m_k))^2 - \sum_t^{T-1}(x_{t+1}(m_k) - \hat{x}_{t+1}(m_k))^2)/(df-1)}{\sum_t^{T-1}(x_{t+1}(m_k) - \hat{x}_{t+1}(m_k))^2/(T-df)}$$

where T is the number of days in the sample and df is the equivalent degrees of freedom for the fit.

4.2.3 Forecast evaluation

We assess the forecast performance of the FDA models using the following measures:

1. Mean absolute error (MAE) is the average of the absolute differences between the forecast, $\hat{x}_{t+1}(m_k)$, and the corresponding observation, $x_{t+1}(m_k)$. It measures the average error magnitude in the forecasts, regardless of error direction and serves to aggregate the errors into a single measure of predictive power.

$$MAE = \frac{1}{T}\sum_{i=1}^{T-1} \left| x_{t+1}(m_k) - \hat{x}_{t+1}(m_k) \right|,$$

   where $x_{t+1}(m_k)$ are the observed values and $\hat{x}_{t+1}(m_k)$ are the values predicted from the model.

2. Root mean squared error (RMSE) is a measure of the difference between values predicted by a model and values realised. The RMSE is defined as the square root of the mean squared error, and again serves to aggregate the errors into a single measure of predictive power.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{T-1}(x_{t+1}(m_k) - \hat{x}_{t+1}(m_k))^2}{T}},$$

   where $x_{t+1}(m_k)$ are the observed values and $\hat{x}_{t+1}(m_k)$ are the values predicted from the model.

3. Mean mixed error (MME) is an asymmetric loss function. MME(U) penalises under-predictions more heavily, while MME(O) penalises over-predictions more heavily. This is very important for investors in option markets, as an under (over)-prediction of implied volatility is more likely to be of greater concern to a seller (buyer) than a buyer (seller). The measure has been employed previously in studies evaluating volatility forecasting techniques such as Brailsford and Faff (1996) and Fuertes et al. (2009).

$$MME(U) = \frac{1}{T}\left[\sum_{t=t_1^O}^{t_N^O} \left| x_{t+1}(m_k) - \hat{x}_{t+1}(m_k) \right| + \sum_{t=t_1^U}^{t_N^U} \sqrt{\left| x_{t+1}(m_k) - \hat{x}_{t+1}(m_k) \right|}\right]$$

and

$$MME(O) = \frac{1}{T}\left[\sum_{t=t_1^O}^{t_N^O}\sqrt{|x_{t+1}(m_k) - \hat{x}_{t+1}(m_k)|} + \sum_{t=t_1^U}^{t_N^U}|x_{t+1}(m_k) - \hat{x}_{t+1}(m_k)|\right],$$

where t_N^U is the number of under-predictions and t_N^O is the number of over-predictions. t_1^O,...,t_N^O represent the indices of the over-predictions, and t_1^U,...,t_N^U represent the indices of the under-predictions.

4. The mean correct predictor of direction of change (MCPDC) is the percentage of predictions for which the forecast, $\hat{x}_{t+1}(m_k)$, has the same sign as the corresponding observation, $x_{t+1}(m_k)$. MCPDC measures how well the model can forecast the direction of movement, regardless of error magnitude. It is also employed in Bernales and Guidolin (2014).

The out-of-sample performance of the FDA-based forecast is benchmarked against traditional models used in the literature:

1. Autoregressive (AR(1)) process of order 1 (Konstantinidi et al. 2008 and Dunis et al. 2013)
2. Generalised autoregressive conditional heteroskedastic (GARCH$(1,1)$) model (Yu et al. 2010 and Dunis et al. 2013). The conditional mean specified is an ARMA(1,1) process, with a normal conditional distribution being assumed.
3. Autoregressive fractionally integrated moving average (ARFIMA$(1,z,1)$) (Konstantinidi et al. 2008). The integrated order of difference is z, where $0 < z < 1$. z is selected using maximum likelihood recursion.

To control for sensitivity to specific out-of-sample periods, various window lengths are tested: 100 day (out-of-sample: July 2013 to November 2013), 200 day (out-of-sample: February 2013 to November 2013), 500 day (out-of-sample: December 2011 to November 2013), and 1000 day (out-of-sample: January 2010 to November 2013). The out-of-sample forecast, between the end of the in-sample period and November 2013, are obtained using a recursive scheme. Each day an additional observation is added to an expanding training window and the models are re-estimated. This is in line with Chalamandaris and Tsekrekos (2010) who adopt a recursive 1-day strategy scheme. Konstantinidi et al. (2008) and Goncalves and Guidolin (2006), also implement out-of-sample recursive schemes by expanding the training window size at 100-day intervals. We choose to expand the training

set and re-estimate the model at each time step, daily, to more incorporate all available up-to-date information into our prediction. This approach more accurately simulates the action likely to be taken by a market practitioner who seeks to predict the following day's movement. The accuracy of these predictions are evaluated using the measures outlined in Section 5.3.3. Subsequently, a formal cross-model comparison is undertaken through the construction of multiple hypothesis tests to ascertain if the FDA-based model produces a more accurate forecast compared to the benchmark models.

4.3 Multiple hypothesis testing

We contribute to the existing literature by incorporating controls for the multiple comparisons problem in our forecasting framework. This robust testing framework adjusts for the likelihood that *seemingly* significant ouperformance can be due to mere chance alone. As we are simultaneously testing, for each out-of-sample window, 300 hypotheses, given 5 forecast evaluation measures, 3 comparative benchmark models, 4 contract maturity lengths, and 5 delta values, the multiple comparisons problem is an issue that must be addressed. The multiple comparison problem states that given enough simultaneous hypothesis tests, statistically significant results may be found by pure chance alone. To control for such false discoveries, the operative balanced stepdown procedure of Romano and Wolf (2010) is employed. The balanced stepdown procedure offers a more generalised and flexible approach to controlling for the multiple comparisons problem than previous frameworks proposed. See Chapter 2 for further motivation. It works by controlling the probability that at least k or more false discoveries occur. Consistent with the notation of Romano and Wolf (2010), the following definition is made for the generalised familywise error rate:

$$k\text{-}FWER_\theta = P_\theta \left\{\text{reject at least } k \text{ null hypothesis } H_{0,s} : s \in \mathcal{I}(\theta)\right\}.$$

$\mathcal{I}(\theta)$ is defined as the set of true null hypotheses and k is a user-defined parameter such that we control for $k \geq 1$ false discoveries. A significance level α is chosen such that k-FWER$\leq \alpha$. The stepdown procedure is constructed such that at each stage, information on the rejected hypotheses to date is used in re-testing for significance on the remaining hypotheses. Within the context of controlling the generalised k-FWER, the overall objective is to ensure that the simultaneous confidence interval covers all parameters except for at most $(k-1)$ of them, for a given limiting probability $(1-\alpha)$, while at the same time ensuring balance (at least asymptotically). Attractive properties of the framework include conservativeness, which allows for finite sample control of the k-FWER under P_θ, and provides asymptotic control in the case of contiguous alternatives.

Towards building a framework to identify outperformance in implied volatility forecast-

ing models, the following hypotheses are considered:

$$H_0 : \theta_{benchmark} - \theta_{FM} \leq 0$$

$$H_1 : \theta_{benchmark} - \theta_{FM} > 0$$

where θ_{FM} is a given forecast evaluation measure for a functional model, and $\theta_{benchmark}$ is the corresponding measure for a given comparative benchmark model. We utilise all five forecast evaluation measures set out in Section 5.3.3, i.e., MAE, RMSE, MME(U), MME(O), and (1-MCPDC). In the latter case, the complement of MCPDC is adopted to conform with the hypothesis setup above. In requiring to circulate through all $(k-1)$-sized subsets of hypotheses rejected up to the current step, to obtain the maximum critical value to apply at each stage of the stepdown procedure, the algorithm can become highly, if not excessively, computationally burdensome. Romano and Wolf (2010) therefore suggest an operative method that reduces this computational burden, while at the same time maintaining much of the attractive properties of the algorithm.[3] It is this operative method that is used for the empirical analysis in subsequent sections.[4]

4.4 Data description

The data set comprises, at-the-money, risk reversal, and butterfly composition implied volatility quotes for the Euro/United States Dollar (EUR-USD) currency pair obtained from Bloomberg. We focus on this single heavily traded currency pair to minimise issues around data quality (i.e., stale and out-of-context quotes). The EUR-USD pair constitutes a developed pair whereby option contracts are the main avenue through which investors exploit the interest rate differentials between the different countries. The use of a contributory data vendor such as Bloomberg, mitigates the idiosyncratic effect specific to individual market participants providing quotes. This issue is cited by Chalamandaris and Tsekrekos (2014), with Bloomberg being used to validate their proprietary J.P. Morgan data set. Through the use of this J.P. Morgan database, Chalamandaris and Tsekrekos (2014) find that implied volatility is more predictable for very liquid currency pairs, citing EUR-USD as an example. EUR-USD is also the sole focus of the study by Dunis et al. (2013). The constant option maturities utilised are: 1, 3, 6 and 9 months. Delta values of 5, 10, 15, 25, 35, 50, 65, 75, 85, 90, 95 are constructed from the at-the-money, risk reversal, and butterfly implied volatility quotes using the Black-Scholes (1973) and Garman and Kohlagen (1983) option pricing formulae. Log changes in implied volatility are calculated for the January 2006 to November 2013 period. As in Chalamandaris and Tsekrekos (2011), we limit our forecast prediction to the surfaces with the highest levels of liquidity. The most liquid contracts are delta values of 10, 25, 50, 75, and 90. It is for this reason that our out-of-sample forecasts concentrate on these particular contracts.

[3] Further technical implementation details can be found in Chapter 2 and Romano and Wolf (2010).

[4] The resampling based MHT algorithms were made available to me by Dr Mark Cummins.

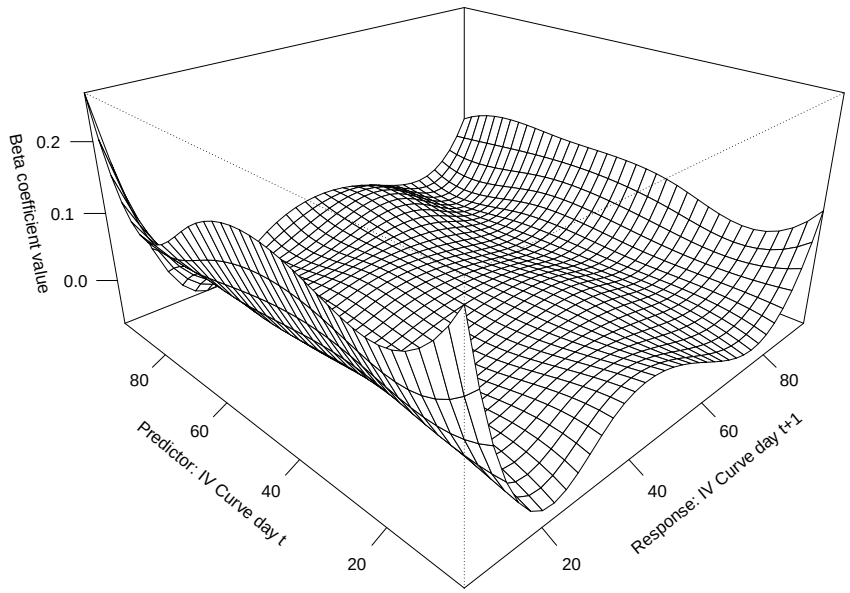
4.5 Empirical results

Firstly, this section presents the results of modelling the evolution of implied volatility using both the fully functional model and the scalar response model for the entire sample, January 2006 to November 2013. The fully functional model is fitted in-sample to ascertain if implied volatility demonstrates intertemporal dependency across the moneyness range. The resultant three-dimensional beta coefficient is plotted and used as an exploratory tool to provide a broad sense of what drives the evolution of implied volatility. Secondly, the results of the scalar response model in-sample fitting are presented with the quality of fit being compared to that of the fully functional model. Thirdly, the out-of-sample forecasts for the scalar response model and the comparative benchmark models only, are evaluated. This is conducted using the measures outlined in Section 5.3.3. Finally, formal testing incorporating the operative balanced stepdown procedure of Romano and Wolf (2010), as set out in Section 5.3.4, is implemented to test if the scalar response model outperforms traditional benchmark models in terms of forecast accuracy.

4.5.1 In-sample functional linear model fit

The fully functional model seen in Equation 5, is employed as an exploratory tool to determine if there is a dependency between implied volatility functions over time. The calculated estimates for the intercept function of the fully functional model are insignificant. For this reason we turn the focus of the analysis to the bivariate regression coefficient function, $\beta(m, m')$, which defines the dependence between the functional predictor and the functional response at each point across the delta range. $\hat{\beta}(m, m')$ estimated from EUR-USD implied volatility for the full sample January 2006 to November 2013 is plotted in Figure 4.1. In line with the success achieved by Konstantinidi et al. (2008) and Dunis et al. (2013), in modelling implied volatility evolution using autoregressive processes, one might intuitively expect the primary driver of the change in the current day's ATM (50 delta) implied volatility to be the change in the previous day's ATM implied volatility. However, it can be seen in Figure 4.1, that the previous day's ATM implied volatility, while important, has less of an impact on the following day's ATM implied volatility than those contracts traded at 20 delta either side of ATM. This gives an empirical indication that the shape and dynamic of the implied volatility function should be incorporated into implied volatility forecasts. The dynamic could be due to non-uniform trading across the curve, as cited by Chalamandaris and Tsekrekos (2010), whereby segments of the implied volatility surface adjust to information at different rates. Ramsay et. al (2009) note that B-spline functions are less stable as they approach their interval boundaries due to less data being available to define their values, with the beginning and end values being determined by only a single coefficient. This feature is evident in Figure 4.1 with large peaks observed at the 5 delta and 95 delta extremities of the response function. To assess model fit and thusly the validity of the inferences drawn, we calculate the R^2 statistic for the fully

Figure 4.1: Fully functional model fitting bivariate regression coefficient

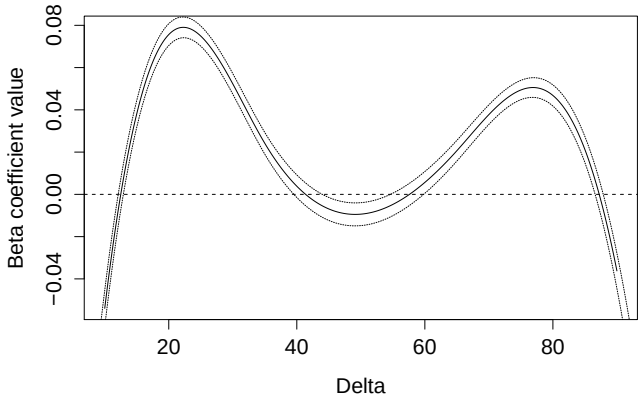


The dependency value uncovered when fitting the fully functional model to the EUR-USD implied volatility data set for the January 2006 to November 2013 period. Option maturity=1 month. "IV" is used as an abbreviation for implied volatility.

functional. R^2 is an informal test that seeks to explain how the models fit the data. The R^2 statistic values for the fully functional model for all deltas and option maturities are given in Table 4.1. The values for the one month maturity contract range from 0.14 to 0.37 across various moneyness levels. Lower R^2 values are observed at in- and out-of-the-money levels, suggesting that the fully functional model provides a comparatively better fit for ATM contracts.

Turning our attention to the scalar response model, it is noted that this model specifies the discrete $t+1$ implied volatilities for each contract as the response, and the implied volatility function on day t as the predictor. As with the fully functional model fitting, intercept values are again insignificant. The estimate of the dependency coefficient, $\hat{\beta}(m)$, is plotted in Figure 4.2. 95% confidence intervals are represented by the dashed lines. Examining $\hat{\beta}(m)$ across the range of moneyness contracts mirrors the results of the fully functional model, whereby the previous day's change observed for contracts traded at in- and out-of-the-money values of delta 20-40 and delta 60-80 range, have a greater impact on the ATM implied volatility level observed today. This parallels the indication of the fully functional model that an FDA-based model may outperform the forecasting performance of a traditionally employed discrete autoregressive process through incorporating information across the entire curve. Analysing the graph also suggests that negative autocorrelation is present, as a positive (negative) previous day ATM implied volatility change is associated with a negative (positive) ATM implied volatility change today. The high R^2 values and

Figure 4.2: Scalar response model fitting regression coefficient



The dependency value uncovered when fitting the scalar response model to the EUR-USD implied volatility data set for the January 2006 to November 2013 period. Delta=50 and option maturity=1 month.

significant F-ratio results calculated in Table 4.2 suggest that the quality of the fit for the scalar response model is far better than one would expect by chance alone.[5]

4.5.2 Out-of-sample forecast evaluation

It is established in the previous section that the scalar response model provides a good fit for modelling the evolution of implied volatility. We now turn our attention to out-of-sample forecasting. A summary of the out-of-sample forecast measures calculated for at-the-money implied volatility under a recursive parameter estimation scheme and a 500 day out-of-sample window length are presented in Table 4.3. The measures calculated for other delta values are given in the Appendix.[6] The results give clear indications that the scalar response model outperforms the traditionally used AR, GARCH and ARFIMA models in forecasting implied volatility out-of-sample. The scalar response model outperforms in terms of both RMSE and MAE across all maturity lengths. The MCPDC results specify that the scalar response model correctly predicts the direction of implied volatility change 88.4%-93.4% of the time. The ARFIMA model is the next best for predicting the direction of change with MCPDC results of 61.4%-79%. The asymmetric mean mixed error loss functions give an indication of which models systematically under- and over- predict implied volatility changes. The closer the MME(U) and MME(O) values for a given model, the lower the level of systematic under- or over- prediction. The MME(U) and MME(O) results presented in Table 4.3 indicate that the scalar response model has a slight tendency to over-predict future implied volatility change. The one month maturity MME(U) and MME(O) values of 0.0298 and 0.0377, respectively, are quite close however, indicating that

[5] In conventional multivariate analysis such high R^2 values could be symptomatic of a problem. However, as documented by Malfait and Ramsay (2003), when a fine evaluation grid is specified, high R^2 follow. For example, R^2 values as high as 99.7% are exhibited in Section 5 of Malfait and Ramsay (2003).

[6] Other out-of-sample window periods, of 100, 200 and 1000 days, are utilised with similar results obtained.

Table 4.1: Fully functional model fitting R^2 statistic values

Delta	R^2 1 month	R^2 3 month	R^2 6 month	R^2 9 month
5	0.273	0.294	0.425	0.370
10	0.180	0.224	0.360	0.264
15	0.145	0.173	0.285	0.213
25	0.207	0.171	0.244	0.246
35	0.310	0.217	0.273	0.295
50	0.373	0.248	0.296	0.322
65	0.300	0.204	0.262	0.283
75	0.197	0.144	0.211	0.212
85	0.140	0.124	0.222	0.158
90	0.176	0.149	0.280	0.184
95	0.262	0.196	0.359	0.269

The R^2 statistic values calculated after fitting the fully functional model to the EUR-USD implied volatility data set over the January 2006 to November 2013 period. Option maturities are 1, 3, 6, and 9 months respectively.

any bias is minor and may be data set specific. The GARCH model produces the most unbiased predictions with one month maturity values of 0.0805 and 0.0846 exhibited for the MME(U) and MME(O) asymmetric loss functions respectively.

As an important contribution to the existing literature, we rigorously evaluate the competing models in an out-of-sample forecast. For this, the operative balanced stepdown procedure of Romano and Wolf (2010) is applied to control for the multiple comparisons problem, as set out in Section 2.3.2. In particular, it controls the generalised FWER using a stepwise procedure that offers balance by construction. This property of balance ensures that each measure is treated equally in terms of power, i.e., the ability to reject false null hypotheses, so measures with large deviations do not dominate those with lower deviations. This is one of the key motivations for using the balanced stepdown procedure for the empirical analysis of this study. To ensure tight control of the number of false discoveries while at the same time offering power to the tests, the generalizing parameter, k, is chosen to ensure that no more than 5% of the 300 tests (per out-of-sample window) represent false discoveries. The probability parameter, α, is set at 5%, such that the probability of $300 \times 5\% = 15$ or more false discoveries is less than or equal to 5%. An N_{max} value of 100 combinations is specified, in line with Romano and Wolf (2010). We specialise the hypotheses set out in Section 5.3.4 as follows:

$$H_0 : \theta_{benchmark} - \theta_{SR} \leq 0$$

$$H_1 : \theta_{benchmark} - \theta_{SR} > 0$$

where θ_{SR} is a given forecast evaluation measure for the scalar response model, and $\theta_{benchmark}$ is the corresponding measure for a given comparative benchmark model.

Table 4.2: Scalar response model (SR) fitting R^2 statistic and F-Ratio values

Delta	R^2 1 Month SR	F-Ratio 1 Month SR	R^2 3 Month SR	F-Ratio 3 Month SR	R^2 6 Month SR	F-Ratio 6 Month SR	R^2 9 Month SR	F-Ratio 9 Month SR
5	0.995	67112.5	0.996	86236.9	0.996	85210.3	0.996	77212.6
10	0.949	6337.1	0.966	9706.0	0.963	8856.4	0.949	6333.5
15	0.949	6394.6	0.978	15093.1	0.986	24710.9	0.974	12947.1
25	0.959	7981.8	0.980	16699.4	0.980	17099.4	0.964	9123.4
35	0.959	7977.6	0.978	15123.4	0.981	17714.1	0.969	10578.7
50	0.951	6563.5	0.967	10080.7	0.963	9000.7	0.944	5770.2
65	0.958	7744.2	0.973	12463.5	0.978	14983.8	0.965	9314.8
75	0.954	7033.4	0.975	13130.4	0.977	14539.8	0.957	7685.6
85	0.940	5328.8	0.969	10791.7	0.980	17159.7	0.965	9424.5
90	0.937	5042.0	0.964	9177.9	0.962	8620.7	0.95	6525.0
95	0.994	58990.6	0.995	62003.2	0.994	58190.9	0.995	63347.7

The R^2 and F-ratio of the scalar response model with option maturity lengths of 1, 3, 6 and 9 months for the January 2006 to November 2013 period. The F-ratio is calculated with 6 degrees of freedom in the numerator and 1548 in the denominator.

Table 4.3: Scalar response (SR), AR, GARCH, and ARFIMA models out-of-sample forecast evaluation measures (Delta=50)

Maturity	1 Month					3 Month				
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0061	0.0047	0.9340	0.0377	0.0298	0.0041	0.0032	0.9100	0.0294	0.0257
AR	0.0326	0.0251	0.4760	0.0884	0.0808	0.0243	0.0182	0.4320	0.0751	0.0644
GARCH	0.0312	0.0240	0.7500	0.0846	0.0805	0.0225	0.0169	0.7320	0.0693	0.0653
ARFIMA	0.0314	0.0241	0.7900	0.0866	0.0792	0.0227	0.0170	0.7680	0.0721	0.0632
Maturity	6 Month					9 Month				
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0039	0.0029	0.9140	0.0276	0.0246	0.0042	0.0031	0.8840	0.0305	0.0229
AR (1)	0.0193	0.0144	0.4520	0.0661	0.0565	0.0170	0.0127	0.5160	0.0595	0.0549
GARCH	0.0175	0.0131	0.7440	0.0596	0.0568	0.0157	0.0118	0.6140	0.0580	0.0524
ARFIMA	0.0186	0.0139	0.7900	0.0645	0.0559	0.0197	0.0148	0.1600	0.0685	0.0558

The forecast evaluation measures calculated after fitting the scalar response model to the EUR-USD implied volatility data set. Delta=50, the in-sample period is January 2006 to December 2011, with the out-of-sample period spanning December 2011 to November 2013.

After applying the Romano and Wolf (2010) procedure, the scalar response model demonstrates truly significant outperformance versus the comparative benchmarks in predicting EUR-USD implied volatility for all 10, 25, 50, 75, and 90 deltas, and all 1, 3, 6, and 9 month option contracts, under the 100, 200, and 500 day out-of-sample window periods. This is concluded for each of the following measures: MAE, RMSE, MCPDC, MME(U), and MME(O). It must be noted however, that the scalar response model, is not shown to significantly outperform the forecast of the AR, ARFIMA, or GARCH models under the RMSE measure, for 50 and 75 delta 9 month maturity contracts, using the 1000 day out-of-sample window. This lack of significance is RMSE-measure specific, as all the alternative forecast evaluation measures calculated for these contracts, MAE, MCPDC, MME(U), and MME(O), are found, in contrast, to be significant. It is only these 6 RMSE tests, from a suite of 1200, that are deemed not to be significant under the Romano and Wolf (2010) framework. Overall, the results clearly indicate that the scalar response model outperforms the traditionally proposed benchmarks in forecasting EUR-USD implied volatility.

4.6 Conclusion

We propose the use of a functional framework to characterise and forecast FX option implied volatility. A major contribution of the study is that of robustly demonstrating the performance advantage of adopting a scalar response model framework to predict future implied volatility movements. The performance of the proposed FDA-based model is benchmarked against the more traditionally employed approaches of the AR, ARFIMA, and GARCH models. FDA boasts the advantages of being computationally efficient, of allowing functions to be evaluated on an arbitrarily fine grid and of removing the need to impose strict parametric structure assumptions. These and other advantages of FDA

are outlined in Ramsay and Silverman (2005). Statistically significant out-of-sample performance is uncovered utilising the adopted measure of MAE, RMSE, MCPDC, MME(U) and MME(O). This is an empirical demonstration that infinite dimensional representation offered by the FDA-based methodology uncovers additional dependencies missed by traditional forecasting models. The findings contradict the conclusion by Dunis et al. (2013) that there is only predictability in the EUR-USD implied volatility process at forecasting horizons of up to 5 hours ahead, and of Chalamandaris & Tsekrekos (2014) who only find predictability at forecast horizons of greater than 5 days. Through the implementation of the fully functional model as an exploratory tool, we conclude that there is intertemporal dependency across the moneyness range. The shape along the implied volatility curve contains important features which should be incorporated to improve the accuracy of forecasts. This is in line with the finding by Chalamandaris and Tsekrekos (2010) who emphasise the need to incorporate the dynamics along the implied volatility function in order to produce accurate forecasts. However, none of their proposed models consistently outperform autoregressive based benchmarks whereas our scalar response model does. Another key finding, is that today's implied volatility function shape can be used to predict the implied volatility level tomorrow, indicating persistence in the process evolution. This study incorporates the novel use of a contributory data vendor to the literature. This mitigates the idiosyncratic risk, highlighted previously by Chalamandaris & Tsekrekos (2014), associated with obtaining quotes from a single market participant.

In a further contribution to the existing literature, a large number of hypotheses are simultaneously tested with robust multiple comparison controls implemented to adjust for false discoveries. This rigorous testing framework concludes that in order to forecast the evolution of implied volatility, the proposed scalar response model provides the greatest level of performance. This paper adds to the growing body of implied volatility modelling literature and could be useful for academics seeking to further understand market efficiency. It has potential pricing implications as the market's expectation of average future volatility between now and option expiry is a major component of many asset pricing models. There may also be potential for speculative traders to exploit the uncovered predictability.[7]

[7]We refrain from presenting an examination of a trading strategy, due to an inability to realistically simulate a live market environment. Previous studies rely on idiosyncratic market assumptions, as well as ignoring implementation issues such as liquidity, strategy drawdowns, margin calls, bid-ask spread considerations, and microstructure effects that might distort any calculated profits.

Chapter 5

Extracting FX forward rate term structure information: Merits of a functional method

5.1 Introduction

Our study adds to the existing literature seeking to extract the informational content of forward foreign exchange rates through the novel proposal of a functional data analysis-based forecasting model. Meese and Rogoff (1983a,b) ascertain that standard exchange rate models do not have the ability to beat forecasts implied by the random walk in the short run. In an attempt to explain this, Engel and West (2005) and Engel et al. (2008) demonstrate that such models imply a near random walk process for the exchange rate, so their power to "beat the random walk" in out-of-sample forecasts is low. Furthermore, it has been demonstrated that the forward rate is not the optimal predictor of future spot rates (Hansen and Hodrick 1980, Frankel 1980, Bilson 1981, Frankel and Rose 1995, and Taylor 1995). Despite this, the question as to whether or not there is information imbedded in forward FX rates persists. Clarida and Taylor (1997) seek to answer this by moving beyond such single-equation methods and conclude that forward premia information is in fact considerable. Their restricted vector error correction model (VECM) constitutes the leading challenger to the seminal work of Meese and Rogoff (1983a,b). The approach is applied in a dynamic out-of-sample forecasting framework resulting in root mean squared error and mean absolute error metrics over 50% lower than those implied by the random walk. The results are confirmed by Clarida et al. (2003) and Sager and Taylor (2014), who establish statistically significant outperformance across different data sets.

We move the problem to an infinite-dimensional space to improve on the forecasting performance achieved by Clarida and Taylor (1997). To this aim, we adopt the scalar response model proposed in Ramsay and Silverman (2005). Specifically, we determine the underlying process that characterises the forward rate term structure, and use its func-

tional principal components, to establish dependency relations between the forward rate term structure and future spot exchange rates. The flexible functional data approach accurately captures the forward rate term structure process, whilst mitigating the need to impose restrictive data structure assumptions on the exchange rate system. A number of applications in the bio-mechanical literature have benefited from functional data frameworks. However, it has only recently been exploited for financial analysis. Muller et al. (2011) for instance study high frequency S&P 500 Index levels, and propose a functional volatility process to model volatility trajectories. Their model shows patterns in volatility and by combining it with prediction techniques and functional regression, it can be used to predict future volatility. Benko et al. (2009) focus on implied volatility, by presenting a new two-sample common factor functional principal component analysis technique and applying it to analyse similarities in stochastic behaviours between implied volatility curves of one- and three- month option contracts on the German-Swiss exchange (EUREX). They highlight the strength of using functional data analysis techniques to characterise market data, an approach which we also adopt.

Claims of superior forecasting performance relative to the random walk outlined in previous foreign exchange forecasting literature are often based on a direct comparison of performance measures without formal tests of statistical significance. What sets the VECM apart is that Clarida et al. (2003) and Sager and Taylor (2014) go beyond a direct comparison of performance measures, by establishing statistically significant outperformance. For comparative purposes with previous studies, we initially present a direct comparison of forecasting performance measures. However, we then apply formal tests to identify instances of statistically significant outperformance for the scalar response model over both the VECM and random walk benchmarks. We first test the hypothesis of forecasting outperformance by implementing a simple t-test of performance measures differentials. In an important extension of the literature we contribute further by incorporating controls for the multiple comparisons problem in testing forecast performance. This robust testing framework adjusts for the likelihood that *seemingly* significant ouperformance can be due to mere chance alone. As the number of simultaneous tests conducted increases, so too does the likelihood of such false discoveries. To solve the multiple comparisons problem issue, we implement the operative balanced stepdown procedure of Romano and Wolf (2010); the first time it has been applied in forecasting the evolution of spot foreign exchange rates.[1] The balanced stepdown procedure offers a generalised and flexible approach to controlling for the multiple comparisons problem. The methodologies used in previous studies raise concerns around the validity of the inferences drawn, insofar as many lack multiple comparisons controls. Our results provide a robust signal of improved forecasting performance relative to the random walk indicating that the forward rate term structure contains statistically significant information about the evolution of the spot exchange rate,

[1]The application of the balanced stepdown procedure of Romano and Wolf (2010) is in line with Cummins and Bucca (2012) and Kearney et al. (2014), who adopt the framework in the identification of profitable statistical arbitrage opportunities and exchange traded fund outperformance, respectively.

above what is embedded in the historic spot rate series itself. Further to this our results provide additional evidence supporting the rejection of the simple risk neutral efficient market hypothesis.

In the next section we provide the theoretical background of the forward rate term structure, while Section 5.3 introduces the scalar response model and Clarida and Taylor (1997) models. Section 5.3.4 details the forecast evaluation framework and the Romano and Wolf (2010) operative balanced stepdown procedure. Section 5.4 presents and discusses the empirical results, with Section 5.5 concluding the paper and drawing implications for future studies.

Summary of contributions

- Can we extract the informational content of forward foreign exchange rates through a functional PCA-based forecasting model?
- How does the performance of the functional PCA-based approach compare with both the random walk and the Clarida and Taylor (1997) VECM?
- Does the forward rate term structure contain information about the evolution of spot exchange rates?

5.2 Risk neutral efficient market hypothesis

A major strength of both our proposed scalar response forecasting model and the Clarida and Taylor (1997) framework, is that they work in spite of the failure of the simple risk neutral efficient market hypothesis (RNEMH) and are agnostic to the precise cause of rejection. RNEMH is predicated on both risk-neutrality and rational expectations, and postulates that the k-period forward rate at time t, f_t^k, is equal to the expectation of the spot rate at time $t+k$, s_{t+k}. This is conditional on information available at time t, Ω_t:

$$0 \equiv f_t^k - E\left(s_{t+k} \mid \Omega_t\right).$$

In other words, it hypothesises that the forward rate is the optimal predictor of the future spot rate. The RNEMH is derived from the combination of two theorems, namely, covered and uncovered interest parity (CIP and UIP, respectively). CIP states that the k-period eurodeposit interest rate differential between the domestic, denoted r^k, and foreign country, denoted $r^{k'}$, is equal to the spot-forward premium, $f_t^k - s_t$:

$$0 \equiv r_t^k - r_t^{k'} - \left(f_t^k - s_t\right).$$

Whereas, UIP is a related no-arbitrage condition that is satisfied without the use of a forward contract. It deems that the interest rate differential is equal to the expected forward rate:

$$0 \equiv r_t^k - r_t^{k'} - E\left(s_{t+k} - s_t \mid \Omega_t\right).$$

Empirically it has been shown that CIP holds (Taylor 1987 and 1989) whereas Chaboud and Wright (2005) show that UIP is rejected at horizons above a few hours, yet Chinn and Merdith (2004) find that UIP cannot be rejected at horizons above five years. Therefore, given the average investor's time horizon it can be taken that UIP does not hold empirically. It follows that the simple RNEMH has been decisively rejected (Hodrick 1987, Froot and Thaler 1990, Taylor 1995, Sarno and Taylor 2002). Various phenomena have been proposed to explain the rejection, including the presence of risk premia (Backus et al. 2001, Farhi and Gabaix 2008, Kellard and Sarantis 2008, Alvarez et al. 2009, and Lustig et al. 2011), inefficient information processing (Froot and Frankel 1989), institutional investor currency flows (Froot and Ramadorai 2005), rational bubbles (Lewis 1989), and the well documented peso problem (Rogoff 1979, Evans and Lewis 1995, Burnside et al. 2011). Our proposed scalar response functional model is contingent on the existence of empirical departures from the RNEMH, therefore it serves as an indirect test for its failure.

5.3 Methodology

This section provides the detail of the functional scalar response model, the Clarida and Taylor (1997) comparative benchmark model and the forecasting evaluation framework employed in the study. It begins by outlining the process of producing a functional representation of the forward rate term structure at each time point. This representation is subsequently forecast using the adopted scalar response model where the infinite-dimensional beta coefficient is specified with a functional principal component basis to solve the underdetermination issue. The theoretical basis of Clarida and Taylor (1997)'s forward premia restrictions and how departures from the RNEMH are accommodated in the VECM framework are given in Subsection 5.3.2. Next, we introduce the performance measures adopted for the empirical analysis section, alongside the formal framework being utilised to evaluate the forecasting performance of each model. Finally, the importance of multiple hypothesis testing controls in the context of our study is highlighted, in conjunction with an overview of Romano and Wolf (2010)'s operative balanced stepdown procedure.

5.3.1 Scalar response model

Functional data analysis (FDA) provides a functional representation of the process underlying a data set. The process is defined over a continuum, where continuum values are generally represented in terms of time or space. In this paper the functions are defined over the domain spanned by the tenors of the forward contracts, k. The function serves to characterise the forward foreign exchange rate process. The FDA methodology has many advantages; it accurately captures the forward rate term structure dynamics, there is no assumed parametric structure, it is computationally efficient, and it results in a process

that can be evaluated on an arbitrarily fine grid.[2]

Using a system of weekly observed spot and forward foreign exchange rates $x\left(k_q\right) := \left\{ f_{t,}^{k_0} f_t^{k_1}, f_t^{k_2}, ..., , f_t^{k_N} \right\}$, we uncover a functional data object, or more simply, the function, denoted $\tilde{x}(k)$, that determines the forward rate term structure dynamics.[3] When constructing a functional data object, a vector of n bases, denoted $\phi_1, ..., \phi_n$, must first be specified. The decision of which basis system to specify is driven by the underlying data's known characteristics. For instance, when modelling periodic data, a Fourier basis expansion, comprised of successive sine/cosine terms, is most commonly applied. However, the forward curve does not exhibit strong cyclical variation, so we choose flexible B-splines for the basis function system. B-spline representation offers a number of strengths, as outlined in de Boor (2001). They are essentially a number of polynomials joined together smoothly at fixed points called knots. The number and positioning of the knots are derived from knowledge of the complexity of the underlying process over particular ranges. We place knots at the discrete forward rate tenors available from the data set, with polynomials describing the tenor interval between the knots. This results in $N+1$ knots: $k_q : k_0 \leq ... \leq k_N$, with the range of the various sub-intervals, $[k_q, k_{q+1}]$, being defined through the placement of these knots. Within each sub-interval, the spline is simply a polynomial of order n. The order is calculated as:

$$order = 1 + degree\, of\, the\, polynomial.$$

Computations with B-splines are extremely efficient as at any one point along the curve they simplify to a polynomial that can be easily evaluated. Adjusting the order of the spline allows for the estimation of derivatives of any degree. In this paper, a second order basis or polygonal, is specified, significantly aiding computational efficiency. Functional structures are approximated as a weighted linear combination of these bases:

$$\tilde{x}(k) \equiv c_1\phi_1(k) + c_2\phi_2(k) + ... + c_n\phi_n(k),$$

where $c_1, ..., c_n$ represent the parameters of the expansion's coefficients. As in Ramsay and Silverman (2005), the coefficients c_j can be chosen by minimising:

$$SSE(c_1, ..., c_n) \equiv \sum_{q=0}^{N} \left[x(k_q) - \tilde{x}(k_q)\right]^2 = \sum_{q=0}^{N} \left[x(k_q) - \sum_{j=1}^{n} c_j\phi_j(k_q) \right]^2$$

where SSE stands for "sum of squared errors". The structure of the obtained functional representation, $\tilde{x}_t(k)$, relies on the assumption that there is a inherent link between consecutive observations along the forward rate tenor curve at a given point in time, t. This is a reasonable assumption that does not in itself constitute a failure of the RNEMH. However, we now proceed by forecasting $\tilde{x}_t(k)$, with the view that the market mechanism

[2]These and other advantages of FDA are outlined in Ramsay and Silverman (2005).

[3]We utilise the representation $f_{t,}^{k_0}$ for the spot rate at time, s_t, for ease of notation.

imparts significant information into the term structure of the forward rates, an exercise dependent on departures from the RNEMH.

A functional linear model is used to predict the evolution of the spot rate. Classical linear models seek to describe the dependency between a response variable and a specified set of predictors. In classical regression, scalar values are used for both the explanatory and response variables. However, in functional linear regression at least one of the observed variables is a curve. Given that the explanatory variable adopted in our study is the forward rate term structure, we employ the use of the scalar response/functional explanatory model ("scalar response" henceforth) of Hovarth and Kokoszka (2012). We utilise the scalar response framework to find the dependency between the current day, t, functional representation of the forward rate term structure, $\tilde{x}_t(k)$, and the k-day ahead, s_{t+k}, scalar response:

$$s_{t+k} = \alpha + \int_{\Omega_k} \beta(k)\tilde{x}_t(k)dk + \varepsilon_t,$$

where Ω_k is the defined forward rate tenor range, and where an estimate, $\hat{\beta}(k)$ is found by minimising:

$$\sum_{t=1}^{T-k} \left(s_{t+k} - \alpha - \int_{\Omega_k} \beta(k)\tilde{x}_t(k)dk \right)^2 .$$

In classical linear regression, there must be fewer explanatory variables than observations. Using a functional explanatory variable, however, acts as an infinite-dimensional predictor of a finite set of responses. This means that an exact fit, leading to $\varepsilon = 0$, is always possible. It also means that an infinite number of possible $\beta(k)$ coefficients produce the same predictions. Dimension reduction through a functional principal component basis representation of $\beta(k)$ can be used to solve this underdetermination issue. In this vein, we now briefly outline the procedure for obtaining functional principal components as proposed by Ramsay and Silverman (2005).

Functional principal component analysis is the search for weight functions, ξ, that correspond to probe scores, ρ_ξ, with the highest possible levels of variation. The probe scores are defined as:

$$\rho_\xi\left(\tilde{x}\left(k\right)\right) \equiv \int \xi\left(k\right)\tilde{x}\left(k\right)dk.$$

As mean is a common mode of variation across functional observations, it is removed, with the residuals, $\tilde{x}\left(k\right) - \bar{\tilde{x}}\left(k\right)$, being probed. The probe score variance,

$$Var\left[\int \xi\left(k\right)\left(\tilde{x}_t\left(k\right) - \bar{\tilde{x}}\left(k\right)\right)dk\right],$$

corresponding to probe weight ξ, is calculated as:

$$max_{\xi}\left\{\sum_{t=1}^{T}\rho_{\xi}^{2}\left(\tilde{x}_{t}\left(k\right)\right)\right\}.$$

A natural size restriction of $\int \xi^{2}(k)\,dk = 1$ is imposed. To ensure that each new principal component function captures a distinct mode of variation, they are required to be orthogonal to those computed previously:

$$\int \xi_{h}(k)\,\xi_{l}(k)\,dk = 0 \quad h = 1, ..., l-1.$$

To construct the functional linear model, we regress the response, s_{t+k}, on the principal components of the constructed forward rate term structure function, $\tilde{x}(k)$. We find that specifying three functional principal components provide a good fit for the forward rate term structure. After absorbing the mean function into the intercept term, we can now define our model, with 0 and 52-week tenor values representing the lower and upper bounds, as:

$$s_{t+k} = \alpha + \int_{0}^{52}\sum_{j=1}^{3}\beta_{j}(k)\,\xi_{j}(k)\,\tilde{x}_{t}(k)dk + \varepsilon_{t}.$$

5.3.2 Clarida and Taylor (1997) VECM

To date, the restricted vector error correction model (VECM) of Clarida and Taylor (1997) is the leading challenger to the seminal work of Meese and Rogoff (1983a,b). For this reason this study adopts the VECM as a comparative benchmark model, alongside the traditionally used random walk. Clarida and Taylor (1997) move beyond single-equation methods and conclude that the information contained in the forward premiums is in fact considerable. The approach is applied in a dynamic recursive out-of-sample forecasting framework that results in root mean squared error and mean absolute error metrics that are up to 50% lower than those implied by the random walk. The framework is also adopted by Clarida et al. (2003) and Sager and Taylor (2014) who confirm the results and demonstrate statistically significant outperformance when applying the model to different data sets. We now outline the theoretical basis for the Clarida and Taylor (1997) approach.

Clarida and Taylor (1997)'s framework shows that, given stationary departures from the RNEMH, γ_t, both spot and forward rate series inherit a common stochastic drift. Based on Beveridge and Nelson (1981) and Stock and Watson (1988), Clarida and Taylor (1997) express the spot exchange rate, s_t, as the sum of two processes:

$$s_{t} = z_{t} + q_{t}, \qquad (1)$$

with z_t representing a random walk with drift and q_t being a zero mean stationary process with finite variance. Clarida and Taylor (1997) then make the assumption that γ_t is $I(0)$, leading to:

$$f_t^k = \gamma_t + k\theta + E_t\left(q_{t+k} \mid \Omega_t\right) + z_t, \qquad (2)$$

where θ is a constant, representing the drift component of the random walk process, z_t. Comparing (1) and (2), we see that both the spot, s_t, and the forward series, f_t^k, share a common stochastic trend, z_t. As defined above, θ, γ_t and $E_t\left(q_{t+k} - q_t \mid \Omega_t\right)$ all constitute $I\left(0\right)$ series. It follows, therefore, that the forward premium, $f_t^k - s_t$, is also stationary, and that the forward and spot rates are cointegrated according to the vector $[1, -1]$:

$$f_t^k - s_t = \gamma_t + k\theta + E_t\left(q_{t+k} - q_t \mid \Omega_t\right). \qquad (3)$$

Given that this is true for any forecasting horizon, k, the cointegrating relationship can be generalised to an $(N+1)$-dimensional system, comprised of the spot and N forward rates, $\left\{s_t, f_t^{k_1}, f_t^{k_2}, ..., , f_t^{k_N}\right\}$. In this case, an N-sized vector encompassing the system's forward premia represent the system's cointegrating equilibria. The strength of the approach is that it identifies both the components and coefficient parameters defining the system's cointegrating space. Consistent with Engle and Granger (1987), a system of spot and N forward rates can be well represented by a vector error correction model (VECM). Therefore, following Clarida and Taylor (1997), we estimate a restricted linear VECM using the maximum likelihood method of Johansen (1991), to obtain 4, 13, 26, and 52-week ahead forecasts of the foreign exchange spot rate[4].

5.3.3 Forecast evaluation

The out-of-sample forecasts for a given horizon k are obtained using a recursive scheme. Each week an additional observation is added to an expanding training window and the models are re-estimated. We choose this testing framework in line with Clarida and Taylor (1997), Clarida et al. (2003) and Sager and Taylor (2014). It ensures that forecasting is conditional only on information available at the time of the forecast, while the weekly expansion and re-estimating procedure serves to incorporate all available up-to-date information into the prediction. The accuracy of the forecasts are evaluated using the following measures:

1. Mean absolute error (MAE) is the average of the absolute differences between the forecast, $\hat{s}_{t+k}$, and the corresponding observation, s_{t+k}. It measures the average error magnitude in the forecasts, regardless of direction and serves to aggregate the errors into a single measure of predictive power.

$$MAE = \frac{1}{T-k}\sum_{i=1}^{T-k}\left|s_{t+k} - \hat{s}_{t+k}\right|,$$

[4]Other alternative VECM estimation techniques, such as the two step methdology detailed in Bredin and Muckley (2011), which attempts to explicitly account for heteroskedasticity in the estimation of rank of the long-run information matrix within the VECM specification could be used. However we adopt Johansen (1991) estimation in line with previous studies.

where s_{t+k} are observed values and $\hat{s}_{t+k}$ are the values predicted from the model.

2. Root mean squared error (RMSE) is a measure of the difference between values predicted by a model and values realised. The RMSE is defined as the square root of the mean squared error, and again serves to aggregate the errors into a single measure of predictive power.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{T-k}(s_{t+k} - \hat{s}_{t+k})^2}{T-k}},$$

where s_{t+k} are observed values and $\hat{s}_{t+k}$ are the values predicted from the model.

We present three different levels of forecast evaluation. Firstly, we assess performance across the models through a direct comparison of forecasting measures. This is in line with the approach of Clarida and Taylor (1997). Secondly, we formally test the hypothesis of outperformance using a standard t-test approach, as outlined in the next section. Thirdly, in an important extension of the existing literature, we employ the use of a resampling-based multiple comparisons testing technique to control for data snooping bias. This forecasting evaluation framework offers robust cross-model comparison, allowing us to ascertain scalar response outperformance relative to both benchmark models, Clarida and Taylor (1997)'s VECM and the notoriously hard-to-beat random walk. The next section details the multiple hypothesis testing technique.

5.3.4 Multiple hypothesis testing

We contribute to the existing literature by incorporating controls for the multiple comparisons problem in our forecasting framework. This robust testing framework adjusts for the likelihood that *seemingly* significant ouperformance can be due to mere chance alone. As we are simultaneously testing 48 hypotheses, given two performance measures, two comparative benchmark models, four forecasting horizons, and three currencies, the multiple comparisons problem is an issue that must be addressed. The problem states that given enough simultaneous hypothesis tests, statistically significant results may be found by pure chance alone. To control for such false discoveries, the operative balanced stepdown procedure of Romano and Wolf (2010) is employed. The balanced stepdown procedure offers a more generalised and flexible approach to controlling for the multiple comparisons problem than previous frameworks proposed. It improves upon formerly proposed single step procedures, by allowing for subsequent iterative steps to identify additional hypothesis rejections and offers balance by construction in the sense that each hypothesis is treated equally in terms of power. It works by controlling the probability that at least k or more

false discoveries occur.[5] Consistent with the notation of Romano and Wolf (2010), the following definition is made for the generalised familywise error rate:

$$k\text{-}FWER_\theta = P_\theta \left\{\text{reject at least } k \text{ null hypothesis } H_{0,s} : s \in \mathcal{I}(\theta)\right\}.$$

$\mathcal{I}(\theta)$ is defined as the set of true null hypotheses and k is a user-defined parameter such that we control for $k \geq 1$ false discoveries. A significance level α is chosen where k-FWER$\leq \alpha$. The stepdown procedure is constructed such that at each stage, information on the rejected hypotheses to date is used in re-testing for significance on the remaining hypotheses. Within the context of controlling the generalised k-FWER, the overall objective is to ensure that the simultaneous confidence interval covers all parameters except for at most $(k-1)$ of them, for a given limiting probability $(1-\alpha)$, while at the same time ensuring balance (at least asymptotically). Attractive properties of the framework include conservativeness, which allows for finite sample control of the k-FWER under P_θ, and provides asymptotic control in the case of contiguous alternatives. Towards building a framework to identify outperformance in the foreign exchange forecasting models, the following hypotheses are considered:

$$H_0 : \theta_{benchmark} - \theta_{SR} \leq 0$$

$$H_1 : \theta_{benchmark} - \theta_{SR} > 0$$

where θ_{SR} is a given forecast evaluation measure for a functional scalar response model, and $\theta_{benchmark}$ is the corresponding measure for the comparative benchmark models; VECM from Clarida and Taylor (1997) and a driftless random walk. We utilise the forecast evaluation performance measures set out in Section 5.3.3, namely MAE and RMSE.

In requiring to circulate through all $(k-1)$-sized subsets of hypotheses rejected up to the current step, to obtain the maximum critical value to apply at each stage of the stepdown procedure, the algorithm can become highly, if not excessively, computationally burdensome. Romano and Wolf (2010) therefore suggest an operative method that reduces this computational burden, while at the same time maintaining much of the attractive properties of the algorithm.[6] It is this operative method that is used for the empirical analysis in subsequent sections.[7]

[5] In an attempt to stay consistent with the notation of Romano and Wolf (2010) we reuse the letter k here. In this context k represents the number of false discoveries in the Romano and Wolf (2010) framework and not the forward tenors as defined in previous sections.

[6] Further technical implementation details can be found in Chapter 2 and Romano and Wolf (2010).

[7] The resampling based MHT algorithms were made available to me by Dr Mark Cummins.

5.4 Data and empirical results

5.4.1 Data

Our data set comprises observations of spot, 4, 13, 26, and 52-week forward rates for Euro, Japanese Yen and British Sterling all versus the U.S. Dollar.[8] Weekly exchange rates are obtained over the period of the 26th week of 1990 (02-Jul-1990) to the 26th week of 2014 (30-Jun-2014), 1253 observations in total for each exchange rate series. Following Sager and Taylor (2014) and Della Corte et al. (2009), our Euro series is proxied by use of the German Deutschmark over the July 1990 to January 1999 period.[9] For the purposes of providing a relative comparison with the results presented previously by Clarida et al. (2003), we designate all but the final three years of the data set as in-sample. The data set is sourced from Thomson Reuters Datastream. The strong theoretical priors outlined in Section 5.3.2 dictate that each currency's forward premia, $f_t^k - s_t$, span the cointegration space according to the vector $[1, -1]$.[10] Therefore we proceed by restricting the basis of the cointegration space through imposing the following condition on the VECM:

$$\beta' x_t = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_t \\ f_t^4 \\ f_t^{13} \\ f_t^{26} \\ f_t^{52} \end{bmatrix}.$$

The VECM is dynamically estimated through the maximum likelihood method of Johansen (1991) to obtain 4, 13, 26, and 52-week ahead forecasts.[11] The sample expands recursively with the optimised VECM being re-estimated at each time step (weekly) as outlined in Section 5.3.3. The out-of-sample forecasting performance of both the scalar response and VECM are outlined in the next section.

5.4.2 Numerical comparison

The goal of the paper is to assess the usefulness of the functional model. To this end the measures RMSE and MAE are adopted to examine out-of-sample forecasting performance. The results presented in Tables 5.1, 5.2 and 5.3, compare the forecasting accuracy of our

[8] We choose the same three currency pairs as Sager and Taylor (2014), who cite that they are the most actively traded pairs according to the Bank for International Settlements (2010).

[9] The use of a weekly data frequency is in line with Clarida and Taylor (1997), Clarida et al. (2003) and Sager and Taylor (2014).

[10] As in Clarida et al. (2003) and Sager and Taylor (2014), we proceed with the restrictions, $[1, -1]$, despite the likelihood ratio test indicating that the null hypothesis of four linearly independent forward premiums comprising the basis for the cointegration space is rejected. Clarida et al. (2003) conclude that although the departures from the precise overidentifying restrictions are statistically significant, they are very small in magnitude.

[11] For further technical VECM estimation details, the reader is directed to Johansen (1991) and Clarida and Taylor (1997). A first-order lag is chosen in line with Clarida and Taylor (1997) who cite algorithmic instability using higher-order lag specifications.

Table 5.1: Results of forecasting exercises: Dollar-Euro

k (weeks)	SR (level)	VECM (ratio)	Random Walk (ratio)
Root mean square error (RMSE)			
4	0.0247	0.971	0.977
13	0.0368	0.936	0.947
26	0.0440	0.949	0.970
52	0.0621	1.090	1.139
Mean absolute error (MAE)			
4	0.0201	0.973	0.979
13	0.0319	0.952	0.972
26	0.0361	0.936	0.949
52	0.0556	1.153	1.213

The performance measure for the functional specification is given in the first column of the table with the second and third columns containing the ratio of the scalar response performance measure to the corresponding VECM and random walk performance measures respectively. Therefore, superior relative performance by the Scalar Response model is indicated by a ratio of less than 1. Forecast period is July 2011 to July 2014. "SR" corresponds to "scalar response" with "VECM" corresponding to "vector error correction model".

Table 5.2: Results of forecasting exercises: Dollar-Sterling

k (weeks)	SR (level)	VECM (ratio)	Random Walk (ratio)
Root mean square error (RMSE)			
4	0.0213	0.993	0.998
13	0.0299	0.992	1.001
26	0.0365	0.960	0.969
52	0.0430	0.825	0.842
Mean absolute error (MAE)			
4	0.0166	0.979	0.982
13	0.0239	0.972	0.983
26	0.0291	0.932	0.941
52	0.0360	0.896	0.912

The performance measure for the functional specification is given in the first column of the table with the second and third columns containing the ratio of the scalar response performance measure to the corresponding VECM and random walk performance measures respectively. Therefore, superior relative performance by the Scalar Response model is indicated by a ratio of less than 1. Forecast period is July 2011 to July 2014. "SR" corresponds to "scalar response" with "VECM" corresponding to "vector error correction model".

Table 5.3: Results of forecasting exercises: Dollar-Yen

k (weeks)	SR (level)	VECM (ratio)	Random Walk (ratio)
Root mean square error (RMSE)			
4	0.0277	0.989	0.994
13	0.0575	0.927	0.936
26	0.0922	0.894	0.905
52	0.1444	0.919	0.935
Mean absolute error (MAE)			
4	0.0214	1.020	1.023
13	0.0439	0.956	0.964
26	0.0667	0.865	0.878
52	0.1261	0.948	0.970

The performance measure for the functional specification is given in the first column of the table with the second and third columns containing the ratio of the scalar response performance measure to the corresponding VECM and random walk performance measures respectively. Therefore, superior relative performance by the Scalar Response model is indicated by a ratio of less than 1. Forecast period is July 2011 to July 2014. "SR" corresponds to "scalar response" with "VECM" corresponding to "vector error correction model".

proposed scalar response model against those of the VECM and random walk alternatives. The performance measure for the functional specification is given in the first column of the table with the second and third columns containing the ratio of the scalar response performance measure to the corresponding VECM and random walk performance measures respectively. Superior relative performance by the scalar response model is indicated by a ratio of less than one. The ratios are calculated for each of the forecasting horizons 4, 13, 26, and 52-week. All forecasts are produced using the same recursive estimation approach. A direct comparison of the performance measures indicate that the scalar response model generally outperforms both the VECM and random walk. This result is broadly similar across all currencies, with the exception of some measure specific underperformance exhibited at the 4-week forecasting horizon for the Japanese Yen. The pockets of under performance exhibited at the 52-week forecasting horizon for the Euro could be attributed to instability in the extrema values the constructed function.[12] Overall, these out-of-sample results are impressive in that they show almost systematic outperformance of the scalar response model over the VECM and random walk approaches. This provides an initial indication of an improvement in the ability of our model to extract useful forecasting information from the term structure of the forward rates over the leading models

[12]We note, as outlined in Section 5.3, that the function characterising the forward rate term structure is defined over the tenor range of 0 weeks to 52 weeks, with the constructed function subsequently being used as an explanatory covariate in the scalar response specification. In constructing the 52-week tenor, there is only one single weekly forward rate data point available to define the coefficients in this range. Such a set-up has been noted to lead to instability in the estimations of the range (see Ramsay and Silverman (2005)), and in turn, as may be the case here, can lead to more volatile forecast predictions in the range. A possible remedy is to use neighbouring forward rates to enrich the data, the 9 month and 2 year tenors for instance, however the use of additional forward rates would be inconsistent with the approach taken in previous studies, against which we aim to provide a relative comparison.

in the literature.

5.4.3 Hypothesis tests

The literature has been split on how best to evaluate forecasting performance; Meese and Rogoff (1983a,b) and Clarida and Taylor (1997) infer model superiority using a direct comparison of performance measure differences such as presented in Section 5.4.2, whilst both Clarida et al. (2003) and Sager and Taylor (2014) statistically test for outperformance. To provide a comparison with these two latter studies we also formally test the hypothesis of outperformance, by implementing a t-test on the performance measure differences. The following hypotheses are considered:

$$H_0 : \theta_{benchmark} - \theta_{SR} \leq 0$$

$$H_1 : \theta_{benchmark} - \theta_{SR} > 0$$

where θ_{SR} is a given forecast evaluation measure (RMSE or MAE) for the functional scalar response model, and $\theta_{benchmark}$ is the corresponding measure for the VECM and random walk benchmarks.

The resulting t-statistics and p-values of the tests are given in Tables 5.4, 5.5, and 5.6. The scalar response model demonstrates statistically significant outperformance among the majority ($27/48$ *instances*) of the 4, 13, 26, and 52-week ahead forecasts across the three currencies. The results are even more impressive when we focus on the Euro and Japanese Yen; finding significant out-of-sample outperformance in $22/32$ hypothesis tests conducted. There are only 5 instances of statistically significant outperformance for the British Pound, however, separate testing concludes that it does not exhibit any instances of either the VECM or random walk models outperforming the scalar response approach.

Given that we are simultaneously testing 48 hypotheses; two performance measures, two comparative benchmark models, four forecasting horizons, and three currencies, the multiple comparisons problem is an issue that must be addressed. As the number of simultaneous tests conducted increases, so too does the likelihood of such false discoveries. Omitting multiple comparisons controls could lead to invalid inferences being drawn. This final step of the analysis goes beyond the approaches taken previously and offers an important extension to the literature through the implementation of the Romano and Wolf (2010) framework. The procedure takes account of the number of simultaneous hypothesis conducted and adjusts for the chances of seemingly significant instances of outperformance. As expected, implementing the MHT framework reduces the number of discoveries, however we still find instances of truly significant outperformance in both the Euro and Japanese Yen. Where the naive t-test finds ten instances of functional outperformance for the Euro, four of these discoveries still hold after implementing the MHT controls. In the case of the Japanese Yen, the 12 identified instances of statistical outperformance under the t-test drops to seven under the Romano and Wolf (2010) framework. The reduction

Table 5.4: Significant outperformance: Dollar-Euro

k (weeks)	SR Vs VECM (t-stats)	SR Vs Random Walk (t-stats)
Root mean square error (RMSE)		
4	**2.950*****,†	**2.355****,†
	(0.002)	(0.010)
13	**3.727*****,†	2.718***
	(0.000)	(0.004)
26	1.789**	0.948
	(0.038)	(0.173)
52	-1.893	-2.757
	(0.969)	(0.997)
Mean absolute error (MAE)		
4	**2.488*****,†	1.789**
	(0.007)	(0.038)
13	2.533***	1.216
	(0.006)	(0.114)
26	1.876**	1.337*
	(0.032)	(0.092)
52	-3.159	-3.962
	(0.999)	(1.000)

"SR Vs VECM" corresponds to statistical outperformance of the scalar response model relative to the Clarida and Taylor (1997) VECM framework for a given performance measure. "SR Vs Random Walk" corresponds to statistical outperformance of the scalar response model relative to a driftless random walk. The calculated t-statistics and p-values of the naive hypothesis test of scalar response model outperformance relative to a benchmark are given. The superscripts *,** , and *** indicate that the hypothesis tests are significant at the 90%, 95% , and 99% levels respectively. The superscript †, is used to represent an instance of truly significant outperformance after applying the resampling based balanced operative stepdown framework of Romano and Wolf (2010). The forecast period is July 2011 to July 2014

Table 5.5: Significant outperformance: Dollar-Sterling

k (weeks)	SR Vs VECM (t-stats)	SR Vs Random Walk (t-stats)
Root mean square error (RMSE)		
4	0.752	0.276
	(0.227)	(0.391)
13	0.292	-0.037
	(0.385)	(0.515)
26	0.856	0.680
	(0.197)	(0.249)
52	2.612***	2.442***
	(0.005)	(0.008)
Mean absolute error (MAE)		
4	1.623*	1.556*
	(0.054)	(0.061)
13	0.831	0.508
	(0.204)	(0.306)
26	1.226	1.090
	(0.112)	(0.139)
52	1.425*	1.244
	(0.079)	(0.108)

"SR Vs VECM" corresponds to statistical outperformance of the scalar response model relative to the Clarida and Taylor (1997) VECM framework for a given performance measure. "SR Vs Random Walk" corresponds to statistical outperformance of the scalar response model relative to a driftless random walk. The calculated t-statistics and p-values of the naive hypothesis test of scalar response model outperformance relative to a benchmark are given. The superscripts *,** , and *** indicate that the hypothesis tests are significant at the 90%, 95% , and 99% levels respectively. The superscript †, is used to represent an instance of truly significant outperformance after applying the resampling based balanced operative stepdown framework of Romano and Wolf (2010). The forecast period is July 2011 to July 2014

Table 5.6: Significant outperformance: Dollar-Yen

k (weeks)	SR Vs VECM (t-stats)	SR Vs Random Walk (t-stats)
Root mean square error (RMSE)		
4	0.897	0.683
	(0.186)	(0.248)
13	**4.628*****,†	**4.468*****,†
	(0.000)	**(0.000)**
26	**6.712** ***,†	**6.388*****,†
	(0.000)	**(0.000)**
52	**5.195** ***,†	4.439***
	(0.000)	(0.000)
Mean absolute error (MAE)		
4	-1.144	-1.640
	(0.872)	(0.948)
13	1.631*	1.466*
	(0.053)	(0.073)
26	**6.045*****,†	**5.727*****,†
	(0.000)	**(0.000)**
52	2.230**	1.324*
	(0.014)	(0.094)

"SR Vs VECM" corresponds to statistical outperformance of the scalar response model relative to the Clarida and Taylor (1997) VECM framework for a given performance measure. "SR Vs Random Walk" corresponds to statistical outperformance of the scalar response model relative to a driftless random walk. The calculated t-statistics and p-values of the naive hypothesis test of scalar response model outperformance relative to a benchmark are given. The superscripts *,** , and *** indicate that the hypothesis tests are significant at the 90%, 95% , and 99% levels respectively. The superscript †, is used to represent an instance of truly significant outperformance after applying the resampling based balanced operative stepdown framework of Romano and Wolf (2010). The forecast period is July 2011 to July 2014

in the number of identified hypothesis rejections is most dramatic however, for the British Pound, whereby none of the 5 significant measures under the t-test are deemed to be true discoveries under the Romano and Wolf (2010) procedure. Despite the truly significant outperformance being confined to just two of the three currency pairs, the results are still encouraging, in that the functional model demonstrates multiple instances of outperformance against the benchmark alternatives of the widely lauded VECM and notoriously hard-to-beat random walk.

5.5 Conclusion

It has been proven that the forward rate is not the optimal predictor of the future spot rate (Hansen and Hodrick 1980, Frankel 1980, Bilson 1981, Frankel and Rose 1995, and Taylor 1995). However, the market mechanism may still impart a significant degree of information into the forward rates. The informational content of the forward rate term structure has been most successfully exploited by Clarida and Taylor (1997) with their dynamic VECM approach predicting spot exchange rates out-of-sample with high precision. Building on this work we offer a novel functional data analysis alternative to exploit the informational content of the forward rates.

While it would be disingenuous to claim that the functional model conclusively beats the VECM across all forecasting horizons, it shows great promise as a forecasting tool. The scalar response model leads to near systematic outperformance in terms of a direct comparison of performance measures, coupled with multiple instances of truly significant outperformance. These favourable functional results are cast in the context of remarkable VECM performance documented in numerous studies to date. The use of the flexible functional framework has the advantage of removing the need to impose prescriptive assumptions on the system of foreign exchange rates. Clarida and Taylor (1997) outline the advantages of moving beyond single-equation methods, whereas this study achieves even greater forecasting performance, by exploiting an infinite-dimensional space representation. The analysis serves to highlight the importance of MHT controls, the absence of which would falsely identify inflated levels of outperformance. This may raise concerns about the validity of inferences drawn in previous studies that do not account for this problem.

The improvement in forecasting performance relative to the random walk indicates that the forward rate term structure contains significant information about the evolution of the spot exchange rate, above what is embedded in the historic spot rate series itself. Further to this, the results reinforce the rejection of the RNEMH. Elliott and Ito (1999) and Dunis and Miao (2007) highlight how even small pockets of predictability can be exploited profitably, with our study providing additional evidence supporting the view that exchange rates are in fact predictable. Therefore, our results further vindicate the use of forward bias currency strategies. In this vein, an assessment of profitability of the scalar response framework in a trading context, may constitute a possible avenue for future research.

Chapter 6

Conclusion

This thesis examines the application of functional data analysis across a number of asset classes, using multiple hypothesis testing techniques to control for data snooping bias. Each of the studies contribute to the literature individually, with the collection emphasising the benefits of adopting both econometric approaches to tackle a wide range of empirical finance problems.

Chapter 2 applies data snooping bias controls to identify ETF pricing deviations. We show that when performance is analysed on a non-risk-adjusted basis only, no ETFs in our sample are identified as displaying any measure of outperformance. It is only the risk-adjusted performance measures that give statistically significant outperformance results. The three key takeaways from the study are, firstly, a high proportion of optimised replication, debt asset class, and global/international ETFs exhibit risk-adjusted premiums, highlighting redemption in kind inefficiencies. Secondly, cross-sector and sectoral funds display similar levels of outperformance. However, energy, precious metals and real estate are industries that beat the market on a risk-adjusted basis. Precious metals became a safe haven for investors due to poor performance in equities over the turbulent 2008-2012 period, with the energy sector being buoyed by increased manufacturing demand from China. The financial services sector, in contrast, registers no market beating funds, primarily due to the credit crisis of 2007-2009 and its legacy. Finally, high expense ratio and recent inception date ETFs are more likely to exhibit index outperformance, which may be of interest to investors seeking to outperform benchmarks. This study succeeds in increasing the understanding of ETF performance alongside providing investors with first-stage guidance in identifying ETFs suitable for their portfolios.

Chapter 3 examines implied volatilty, jump risk and pricing dynamics in crude oil markets using functional representation. We find strong evidence of converse jump dynamics in crude oil markets during periods of demand and supply side weakness. The entire set of price dynamics within crude oil markets cannot be fully represented by traditional multivariate analysis so we combine the application of FDA techniques, jump components from the widely implemented Merton model, and an analysis of underlying economic pressures to better understand market implied volatility and jump dynamics in the 2007-2013 sample

period. This is used as a basis for a functional data analysis-derived Merton (1976) jump diffusion optimised delta hedging strategy, which exhibits superior portfolio management results over traditional methods. We make several contributions to the discussion of the relationship between volatility smile slope and curvature dynamics and the parameters of the Merton model. We achieve this by combining the resultant functional data object with key attributes of the Merton model to derive implied values for the average jump amplitude in a manner similar to Yan (2011). We can clearly ex-post demonstrate the link between these values and contemporary socio-economic events, especially during the turbulent years we examine. In our sample, the systematic analysis of implied volatility also highlights periods of economic weakness in advance of their occurrence. Our FDA optimised Merton delta hedging strategy outperforms the Black-Scholes delta hedging strategy by 8% in terms of implementation cost, over the entire sample. Breaking the sample down into periods split by predominantly positive and predominantly negative implied volatility slopes, we see that the Merton strategy outperforms the Black-Scholes when $\hat{k}$ values are positive and broadly matches its performance in periods of negative $\hat{k}$ values.

Chapter 4 robustly demonstrates the performance advantage of adopting a scalar response model framework to predict future implied volatility movements in FX markets. The performance of the proposed FDA-based model is benchmarked against traditionally employed approaches of Gonclaves and Guidolin (2006) and Konstantinidi et al. (2008). The study constitutes an empirical demonstration that infinite dimensional representation uncovers additional dependencies missed by traditional forecasting models. Through the implementation of the fully functional model as an exploratory tool, we conclude that there is intertemporal dependency across the moneyness range. The shape along the implied volatility curve contains important features that should be incorporated to improve the accuracy of forecasts. This is in line with the finding by Chalamandaris and Tsekrekos (2010) who emphasise the need to incorporate the dynamics along the implied volatility function in order to produce accurate forecasts. Another key finding, is that today's implied volatility function shape can be used to predict the implied volatility level tomorrow, indicating persistence in the evolution of the process. Statistically significant out-of-sample performance is uncovered utilising the adopted performance measures. The findings contradict the conclusion by Dunis et al. (2013) that there is only predictability in the EUR-USD implied volatility process at forecasting horizons of up to 5 hours ahead, and of Chalamandaris & Tsekrekos (2014) who only find predictability at forecast horizons of greater than 5 days.

Chapter 5 forecasts spot foreign exchange rates using a functional model to exploit the information contained in currency forwards. We find that the scalar response model leads to near systematic outperformance in terms of a direct comparison of performance measures, verus both the VECM and RW, coupled with multiple instances of truly significant outperformance. These favourable functional results are cast in the context of remarkable VECM performance documented in numerous studies to date. The use of the flexible functional framework has the advantage of removing the need to impose prescriptive as-

sumptions on the system of foreign exchange rates. The analysis also serves to highlight the importance of multiple hypothesis testing controls, the absence of which would falsely identify inflated levels of outperformance. The improvement in forecasting performance relative to the random walk indicates that the forward rate term structure contains significant information about the evolution of the spot exchange rate, above what is embedded in the historic spot rate series itself. Further to this, the results reinforce the rejection of the RNEMH. Elliott and Ito (1999) and Dunis and Miao (2007) highlight how even small pockets of predictability can be exploited profitably, with our study providing additional evidence supporting the view that exchange rates are in fact predictable. As a result, our results further vindicate the use of forward bias currency strategies.

Overall, exploiting the infinite dimensional representation offered by functional data techniques are shown to result in forecasting and hedging performance benefits. This conclusion is reinforced through the application of the MHT framework, as economic arguments are intrinsically linked to the robustness of the econometric analysis. We demonstrate forecasting success in Chapters 4 and 5 using a functional data framework that its robust to generalised correction for data snooping bias. Therefore, it can be concluded that the implementation of a joint FDA and MHT approach constitutes a powerful empirical finance forecasting tool. Furthermore, the techniques outlined here are of benefit to the wider investment community as an aid in identifying specific investments suitable to individual portfolio requirements. Potential limitations of the thesis centre on data availability, as ideally we would have access to intraday ETF quotes, given that additional pricing deviations may only hold at a higher frequency than our daily data indicates. Despite having not fully exhausted the application of FDA and MHT procedures in respect to the questions asked in this thesis, a separate strand of research could look at the application of functional data techniques to forecast yield curve evolution. Furthermore, evaluating efficiency relations in foreign exchange markets represents another potential avenue for future research.

References

Ackert, L., and Y. Tian. 2008. Arbitrage, liquidity, and the valuation of exchange traded funds. *Financial Markets, Institutions & Instruments* 17:331–362.

Alexander, C., and A. Barbosa. 2008. Hedging index exchange traded funds. *Journal of Banking & Finance* 32:326–337.

Alvarez, F., A. Atkeson, and P. J. Kehoe. 2009. Time-varying risk, interest rates, and exchange rates in general equilibrium. *The Review of Economic Studies* 76:851–878.

Antonio, F. D. Narzo, J. L. Aznarte, and M. Stigler. 2009. *tsDyn: Time series analysis based on dynamical systems theory*. R package version 0.7.

Askari, H., and N. Krichene. 2008. Oil price dynamics (2002–2006). *Energy Economics* 30:2134–2153.

Backus, D. K., S. Foresi, and C. I. Telmer. 2001. Affine term structure models and the forward premium anomaly. *The Journal of Finance* 56:279–304.

Bajgrowicz, P., and O. Scaillet. 2012. Technical trading revisited: false discoveries, persistence tests, and transaction costs. *Journal of Financial Economics* 106:473–491.

Bakshi, G., C. Cao, and Z. Chen. 1997. Empirical performance of alternative option pricing models. *The Journal of Finance* 52:2003–2049.

Ball, C. A., and W. N. Torous. 1983. A simplified jump process for common stock returns. *Journal of Financial And Quantitative Analysis* 18:53–65.

Bannouh, K., M. Martens, and D. van Dijk. 2013. Forecasting volatility with the realized range in the presence of noise and non-trading. *The North American Journal of Economics and Finance* 26:535–551.

Barras, L., O. Scaillet, and R. Wermers. 2010. False discoveries in mutual fund performance: measuring luck in estimated alphas. *Journal of Finance* 65:179–216.

Bates, D. S. 1991. The crash of '87: Was it expected? The evidence from options markets. *The Journal of Finance* 46:1009–1044.

Bates, D. S. 1996. Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *Review of Financial Studies* 9:69–107.

Beber, A., F. Breedon, and A. Buraschi. 2010. Differences in beliefs and currency risk premiums. *Journal of Financial Economics* 98:415–438.

Benko, M., W. Härdle, and A. Kneip. 2009. Common functional principal components. *The Annals of Statistics* 37:1–34.

Bernales, A., and M. Guidolin. 2014. Can we forecast the implied volatility surface dynamics for CBOE equity options? *Journal of Banking & Finance* p. Forthcoming.

Bilson, J. F. 1981. The "speculative efficiency" hypothesis. *Journal of Business* pp. 435–451.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81:637–654.

Borensztein, E. R., and M. P. Dooley. 1987. Options on foreign exchange and exchange rate expectations. *IMF Staff Papers* 34:643–680.

Brailsford, T. J., and R. W. Faff. 1996. An evaluation of volatility forecasting techniques. *Journal of Banking & Finance* 20:419–438.

Bredin, D., and C. Muckley. 2011. An emerging equilibrium in the EU emissions trading scheme. *Energy Economics* 33:353–362.

Burnside, C., M. Eichenbaum, I. Kleshchelski, and S. Rebelo. 2011. Do peso problems explain the returns to the carry trade? *Review of Financial Studies* 24:853–891.

Chaboud, A. P., and J. H. Wright. 2005. Uncovered interest parity: It works, but not for long. *Journal of International Economics* 66:349–362.

Chalamandaris, G., and A. E. Tsekrekos. 2010. Predictable dynamics in implied volatility surfaces from OTC currency options. *Journal of Banking & Finance* 34:1175–1188.

Chalamandaris, G., and A. E. Tsekrekos. 2014. Predictability in implied volatility surfaces: evidence from the Euro OTC FX market. *The European Journal of Finance* 20:33–58.

Chang, C.-L., M. McAleer, and R. Tansuchat. 2013. Conditional correlations and volatility spillovers between crude oil and stock index returns. *The North American Journal of Economics and Finance* 25:116–138.

Chinn, M. D., and G. Meredith. 2004. Monetary policy and long-horizon uncovered interest parity. *IMF Staff Papers* pp. 409–430.

Chuang, W.-I., T.-C. Huang, and B.-H. Lin. 2013. Predicting volatility using the Markov-switching multifractal model: Evidence from S&P 100 index and equity options. *The North American Journal of Economics and Finance* 25:168–187.

Clarida, R. H., L. Sarno, M. P. Taylor, and G. Valente. 2003. The out-of-sample success of term structure models as exchange rate predictors: A step beyond. *Journal of International Economics* 60:61–83.

Clarida, R. H., and M. P. Taylor. 1997. The term structure of forward exchange premiums and the forecastability of spot exchange rates: Correcting the errors. *Review of Economics and Statistics* 79:353–361.

Cont, R., and P. Tankov. 2004. Non-parametric calibration of jump-diffusion option pricing models. *Journal of Computational Finance* 7:1–50.

Corrado, C. J., and T. W. Miller. 2006. Estimating expected excess returns using historical and option-implied volatility. *Journal of Financial Research* 29:95–112.

Craven, P., and G. Wahba. 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* 31:377–403.

Criton, G., and O. Scaillet. 2011. Time-varying analysis in risk and hedge fund performance: how forecast ability increases estimated alpha.

Cummins, M., and A. Bucca. 2012. Quantitative spread trading on crude oil and refined products markets. *Quantitative Finance* 12:1857–1875.

Curcio, R. J., J. M. Lipka, and J. H. J. Thornton. 2004. Cubes and individual investors. *Financial Services Review* 13:123–139.

Cuthbertson, K., D. Nitzsche, and N. O. Sullivan. 2008. UK mutual fund performance: skill or luck? *Journal of Empirical Finance* 15:613–634.

De Boor, C. 2001. *A Practical Guide to Splines*. Springer-Verlag.

DeFusco, R., S. Ivanov, and G. Karels. 2011. The exchange traded funds pricing deviation: analysis and forecasts. *Journal of Economics and Finance* 35:181–197.

Della Corte, P., L. Sarno, and I. Tsiakas. 2009. An economic evaluation of empirical exchange rate models. *Review of Financial Studies* 22:3491–3530.

Dunis, C., N. M. Kellard, and S. Snaith. 2013. Forecasting EUR–USD implied volatility: The case of intraday data. *Journal of Banking & Finance* 37:4943–4957.

Dunis, C. L., and J. Miao. 2007. Trading foreign exchange portfolios with volatility filters: The carry model revisited. *Applied Financial Economics* 17:249–255.

Elliott, G., and T. Ito. 1999. Heterogeneous expectations and tests of efficiency in the Yen/Dollar forward exchange rate market. *Journal of Monetary Economics* 43:435–456.

Elton, E. J., M. J. Gruber, G. Comer, and K. Li. 2002. Spiders: Where are the bugs? *Journal of Business* 75:453–472.

Engel, C., N. C. Mark, and K. D. West. 2008. Exchange Rate Models Are Not As Bad As You Think. In *NBER Macroeconomics Annual 2007, Volume 22*, pp. 381–441. University of Chicago Press.

Engel, C., and K. D. West. 2005. Exchange Rates and Fundamentals. *Journal of Political Economy* 113:485–517.

Engle, R., and D. Sarkar. 2006. Premiums-discounts and exchange traded funds. *Journal of Derivatives* 13:27–45.

Engle, R. F., and B. S. Yoo. 1987. Forecasting and testing in co-integrated systems. *Journal of Econometrics* 35:143–159.

Evans, M. D., and K. K. Lewis. 1995. Do long-term swings in the dollar affect estimates of the risk premia? *Review of Financial Studies* 8:709–742.

Fama, E. F. 1965. The behavior of stock-market prices. *The Journal of Business* 38:34–105.

Farhi, E., and X. Gabaix. 2008. Rare disasters and exchange rates. Tech. rep., National Bureau of Economic Research.

Frankel, J. A. 1980. Tests of rational expectations in the forward exchange market. *Southern Economic Journal* pp. 1083–1101.

Frankel, J. A., and A. K. Rose. 1995. Empirical research on nominal exchange rates. *Handbook of International Economics* 3:1689–1729.

Froot, K. A., and J. A. Frankel. 1989. Forward discount bias: Is it an exchange risk premium? *The Quarterly Journal of Economics* pp. 139–161.

Froot, K. A., and T. Ramadorai. 2005. Currency Returns, Intrinsic Value, and Institutional-Investor Flows. *The Journal of Finance* 60:1535–1566.

Fuertes, A.-M., M. Izzeldin, and E. Kalotychou. 2009. On forecasting daily stock volatility: the role of intraday information and market conditions. *International Journal of Forecasting* 25:259–281.

Garman, M. B., and S. W. Kohlhagen. 1983. Foreign currency option values. *Journal of International Money and Finance* 2:231–237.

Garvey, J. F., and L. A. Gallagher. 2012. The realised–implied volatility relationship: Recent empirical evidence from FTSE-100 stocks. *Journal of Forecasting* 31:639–660.

Gastineau, G. 2001. Exchange-traded funds: an introduction. *Journal of Portfolio Management* 27:88–96.

Gastineau, G. 2004. The benchmark index ETF performance problem. *The Journal of Portfolio Managemen* 30:96–103.

Giot, P., S. Laurent, and M. Petitjean. 2010. Trading activity, realized volatility and jumps. *Journal of Empirical Finance* 17:168–175.

Goncalves, S., and M. Guidolin. 2006. Predictable dynamics in the S&P 500 index options implied volatility surface*. *The Journal of Business* 79:1591–1635.

Gruber, M. J. 1996. Another puzzle: the growth in actively managed mutual funds. *Journal of Finance* 55:783–810.

Hammoudeh, S., P. Araújo Santos, and A. Al-Hassan. 2013. Downside risk management and VaR-based optimal portfolios for precious metals, oil and stocks. *The North American Journal of Economics and Finance* 25:318–334.

Hansen, L. P., and R. J. Hodrick. 1980. Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *The Journal of Political Economy* pp. 829–853.

Hansen, P. R. 2005. A test for superior predictive ability. *Journal of Business & Economic Statistics* 23:365–380.

Harper, J. T., J. Madura, and O. Schnusenberg. 2006. Performance comparison between exchange-traded funds and closed-end country funds. *Journal of International Financial Markets, Institutions and Money* 16:104–122.

Horváth, L., and P. Kokoszka. 2012. *Inference for functional data with applications*. Springer.

Hsu, P.-H., Y.-C. Hsu, and C.-M. Kuan. 2010. Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance* 17:471–484.

Hsu, P.-H., and C.-M. Kuan. 2005. Re-examining the profitability of technical analysis with reality check.

Jares, T., and A. Lavin. 2004. Japan and Hong Kong exchange-traded funds (ETFs): discounts, returns, and trading strategies. *Journal of Financial Services Research* 25:57–66.

Jorion, P. 1988. On jump processes in the foreign exchange and stock markets. *Review of Financial Studies* 1:427–445.

Kearney, F., M. Cummins, and F. Murphy. 2014. Outperformance in exchange-traded fund pricing deviations: Generalized control of data snooping bias. *Journal of Financial Markets* 19:86–109.

Kellard, N., and N. Sarantis. 2008. Can exchange rate volatility explain persistence in the forward premium? *Journal of Empirical Finance* 15:714–728.

Kneip, A., and K. J. Utikal. 2001. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* 96:519–542.

Konstantinidi, E., G. Skiadopoulos, and E. Tzagkaraki. 2008. Can the evolution of implied volatility be forecasted? Evidence from European and US implied volatility indices. *Journal of Banking & Finance* 32:2401–2411.

Lewis, K. K. 1989. Changing beliefs and systematic rational forecast errors with evidence from foreign exchange. *The American Economic Review* pp. 621–636.

Liu, C., S. Ray, G. Hooker, and M. Friedl. 2012. Functional factor analysis for periodic remote sensing data. *The Annals of Applied Statistics* 6:601–624.

Liu, P., and K. Tang. 2011. The stochastic behavior of commodity prices with heteroskedasticity in the convenience yield. *Journal of Empirical Finance* 18:211–224.

Lustig, H., N. Roussanov, and A. Verdelhan. 2011. Common risk factors in currency markets. *Review of Financial Studies* 24:3731–3777.

Malfait, N., and J. Ramsay. 2003. The historical functional linear model. *Canadian Journal of Statistics* 31:115–128.

Malkiel, B. 1995. Returns from investing in equity mutual funds 1971 to 1991. *Journal of Finance* 50:549–572.

Mandelbrot, B. 1967. The variation of some other speculative prices. *The Journal of Business* 40:393–413.

Marshall, B. R., R. H. Cahan, and J. M. Cahan. 2008. Can commodity futures be profitably traded with quantitative market timing strategies? *Journal of Banking & Finance* 32:1810–1819.

Mateus, C., and T. Kuo. 2008. The performance and persistence of exchange-traded funds: evidence for iShares MSCI country-specific ETFs. In *Swiss Society for Financial Market Research 11th Conference.*

Meese, R., and K. Rogoff. 1983a. The out-of-sample failure of empirical exchange rate models: Sampling error or misspecification? In *Exchange Rates and International Macroeconomics*, pp. 67–112. University of Chicago Press.

Meese, R. A., and K. Rogoff. 1983b. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics* 14:3–24.

Merton, R. C. 1976. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3:125–144.

Moschini, G., and R. J. Myers. 2002. Testing for constant hedge ratios in commodity markets: a multivariate GARCH approach. *Journal of Empirical Finance* 9:589–603.

Müller, H.-G., R. Sen, and U. Stadtmüller. 2011. Functional data analysis for volatility. *Journal of Econometrics* 165:233–245.

Murphy, F., and E. Ronn. 2015. The value and information content of options on crude-oil futures contracts. *Review of Derivatives Research* Forthcoming.

Muzzioli, S. 2010. Option-based forecasts of volatility: an empirical study in the DAX-index options market. *The European Journal of Finance* 16:561–586.

Nomikos, N. K., and O. A. Soldatos. 2010. Analysis of model implied volatility for jump diffusion models: Empirical evidence from the Nordpool market. *Energy Economics* 32:302–312.

Park, C.-H., and S. H. Irwin. 2007. What do we know about the profitability of technical analysis? *Journal of Economic Surveys* 21:786–826.

Phengpis, C., and P. E. Swanson. 2009. iShares and the U.S. market risk exposure. *Journal of Business Finance and Accounting* 36:972–986.

Plantinga, A., R. van der Meer, and F. Sortino. 2001. The impact of downside risk on risk-adjusted performance of mutual funds in the Euronext markets. *Geneva Papers on Risk and Insurance* pp. 1–14.

Qui, M., and Y. Wu. 2008. Technical trading-rule profitability, data snooping, and reality check: evidence from the foreign exchange market. *Journal of Money, Credit and Banking* 38:2135–2158.

Ramsay, J., G. Hooker, and S. Graves. 2009. *Functional data analysis with R and MATLAB*. Springer.

Ramsay, J., and B. Silverman. 2005. *Functional data analysis*. Springer, New York.

Ramsay, J., H. Wickham, S. Graves, and G. Hooker. 2011. fda: Functional data analysis. *R package version* 2.

Rogoff, K. 1979. *Essays on expectations and exchange rate dynamics*. Ph.D. thesis, MIT.

Romano, J. P., and A. M. Shaikh. 2006. On stepdown control of the false discovery proportion. *Lecture Notes-Monograph Series* 2:33–50.

Romano, J. P., A. M. Shaikh, and M. Wolf. 2010. Hypothesis testing in econometrics. *Annual Review of Economics* 2:75–104.

Romano, J. P., and M. Wolf. 2007. Control of generalized error rates in multiple testing. *Annals of Statistics* 35:1378–1408.

Romano, J. P., and M. Wolf. 2010. Balanced control of generalized error rates. *The Annals of Statistics* 38:598–633.

Rompotis, G. G. 2011. Predictable patterns in ETFs return and tracking error. *Studies in Economics and Finance* 28:14–35.

Sabbaghi, O. 2011. The behavior of green exchange-traded funds. *Managerial Finance* 37:426–441.

Sager, M., and M. P. Taylor. 2014. Generating currency trading rules from the term structure of forward foreign exchange premia. *Journal of International Money and Finance* 44:230–250.

Sarno, L., and M. P. Taylor. 2002. *The economics of exchange rates*. Cambridge University Press.

Sharpe, W. F. 1966. Mutual fund performance. *The Journal of Business* 39:119–138.

Sullivan, R., and A. Timmermann. 1999. Data-snooping, technical trading rule performance and the bootstrap. *Journal of Finance* 54:1647–1691.

Taylor, M. P. 1987. Covered interest parity: A high-frequency, high-quality data study. *Economica* pp. 429–438.

Taylor, M. P. 1989. Covered interest arbitrage and market turbulence. *The Economic Journal* pp. 376–391.

Taylor, M. P. 1995. The economics of exchange rates. *Journal of Economic Literature* pp. 13–47.

Taylor, S. J., P. K. Yadav, and Y. Zhang. 2010. The information content of implied volatilities and model-free volatility expectations: Evidence from options written on individual stocks. *Journal of Banking & Finance* 34:871–881.

Trolle, A. B., and E. S. Schwartz. 2009. Unspanned stochastic volatility and the pricing of commodity derivatives. *Review of Financial Studies* 22:4423–4461.

White, H. 2000. A reality check for data snooping. *Econometrica* 68:1097–1126.

Wilmot, N. A., and C. F. Mason. 2013. Jump processes in the market for crude oil. *Energy Journal* 34:33–48.

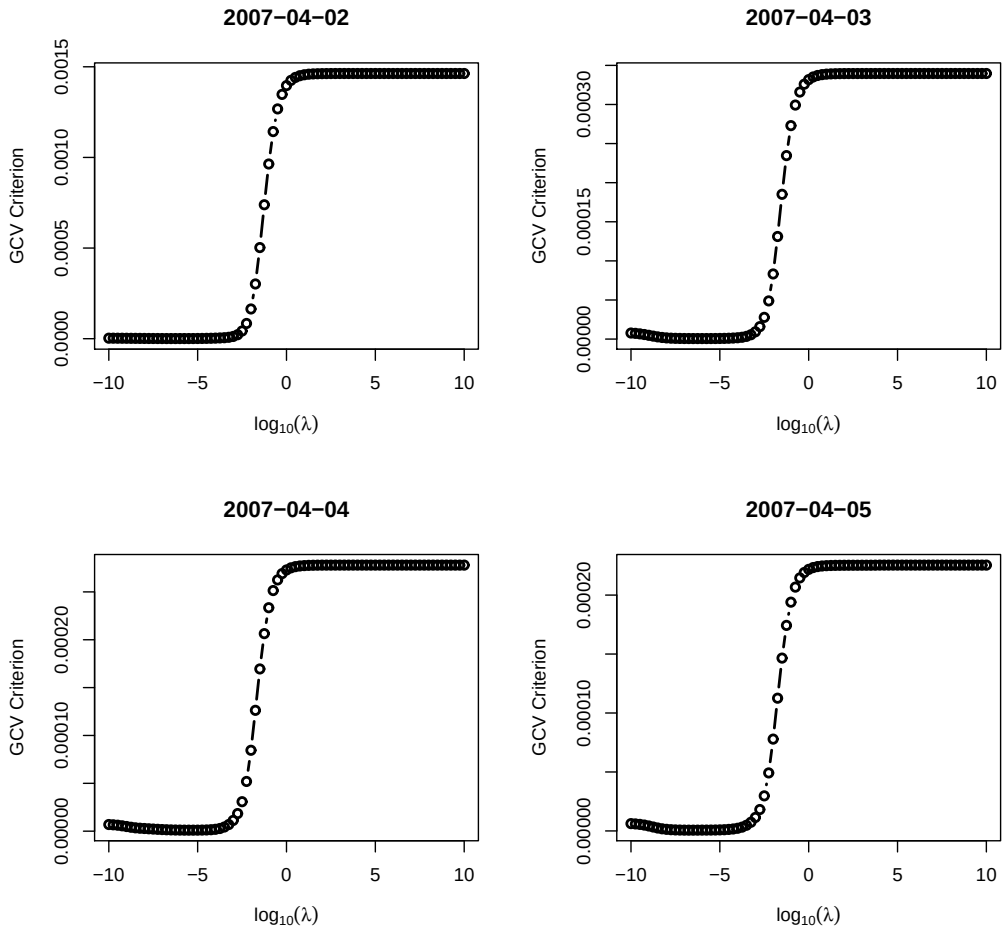
Yan, S. 2011. Jump risk, stock returns, and slope of implied volatility smile. *Journal of Financial Economics* 99:216–233.

Yu, L. 2005. Basket securities, price formation, and informational efficiency.

Yu, W. W., E. C. Lui, and J. W. Wang. 2010. The predictive power of the implied volatility of options traded OTC and on exchanges. *Journal of Banking & Finance* 34:1–11.

Appendix A

Figure 1: Generalised cross validation results



The values of the GCV criterion for choosing the optimal smoothing parameter, λ^*, when constructing implied volatility curve functional data objects. The plots are for the first four days of the sample.

Table 1: Scalar response (SR), AR, GARCH, and ARFIMA models out-of-sample forecast evaluation measures (Delta=10)

Maturity			1 Month					3 Month		
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0063	0.0048	0.9540	0.0374	0.0300	0.0045	0.0034	0.9280	0.0310	0.0249
AR	0.0358	0.0274	0.4360	0.0931	0.0855	0.0271	0.0199	0.4380	0.0799	0.0671
GARCH	0.0375	0.0287	0.3940	0.0949	0.0887	0.0258	0.0189	0.7660	0.0751	0.0674
ARFIMA	0.0366	0.0281	0.5380	0.0951	0.0860	0.0257	0.0189	0.7960	0.0772	0.0653
Maturity			6 Month					9 Month		
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0040	0.0029	0.9380	0.0279	0.0230	0.0039	0.0027	0.9320	0.0281	0.0206
AR (1)	0.0211	0.0159	0.4380	0.0710	0.0589	0.0180	0.0137	0.5200	0.0626	0.0570
GARCH	0.0205	0.0153	0.6520	0.0667	0.0600	0.0180	0.0137	0.5280	0.0618	0.0579
ARFIMA	0.0212	0.0160	0.5220	0.0716	0.0590	0.0201	0.0153	0.0800	0.0688	0.0583

The forecast evaluation measures calculated after fitting the scalar response model to the EUR-USD implied volatility data set. Delta=10, the in-sample period is January 2006 to December 2011, with the out-of-sample period spanning December 2011 to November 2013.

Table 2: Scalar response (SR), AR, GARCH, and ARFIMA models out-of-sample forecast evaluation measures (Delta=25)

Maturity			1 Month					3 Month		
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0043	0.0033	0.9740	0.0318	0.0242	0.0024	0.0018	0.9680	0.0226	0.0177
AR	0.0338	0.0259	0.4620	0.0909	0.0823	0.0255	0.0190	0.4560	0.0777	0.0663
GARCH	0.0325	0.0249	0.7560	0.0869	0.0819	0.0239	0.0179	0.7720	0.0722	0.0668
ARFIMA	0.0325	0.0248	0.8100	0.0884	0.0803	0.0240	0.0179	0.8020	0.0747	0.0646
Maturity			6 Month					9 Month		
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0021	0.0015	0.9680	0.0199	0.0163	0.0021	0.0016	0.9500	0.0227	0.0153
AR (1)	0.0199	0.0148	0.4420	0.0683	0.0565	0.0174	0.0128	0.5320	0.0599	0.0544
GARCH	0.0187	0.0139	0.7120	0.0634	0.0575	0.0168	0.0124	0.5980	0.0586	0.0538
ARFIMA	0.0193	0.0144	0.7300	0.0670	0.0558	0.0198	0.0146	0.1560	0.0674	0.0557

The forecast evaluation measures calculated after fitting the scalar response model to the EUR-USD implied volatility data set. Delta=25, the in-sample period is January 2006 to December 2011, with the out-of-sample period spanning December 2011 to November 2013.

Table 3: Scalar response (SR), AR, GARCH, and ARFIMA models out-of-sample forecast evaluation measures (Delta=75)

Maturity			1 Month					3 Month		
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0046	0.0036	0.9800	0.0330	0.0255	0.0025	0.0019	0.9680	0.0220	0.0193
AR	0.0320	0.0246	0.4860	0.0873	0.0802	0.0241	0.0182	0.4640	0.0736	0.0664
GARCH	0.0310	0.0238	0.7020	0.0842	0.0799	0.0223	0.0169	0.7720	0.0684	0.0663
ARFIMA	0.0316	0.0243	0.6660	0.0867	0.0796	0.0228	0.0172	0.8000	0.0709	0.0647
Maturity			6 Month					9 Month		
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0023	0.0016	0.9580	0.0200	0.0172	0.0023	0.0018	0.9700	0.0225	0.0172
AR (1)	0.0200	0.0152	0.4740	0.0669	0.0605	0.0169	0.0127	0.5260	0.0592	0.0556
GARCH	0.0192	0.0146	0.6460	0.0629	0.0616	0.0154	0.0116	0.6260	0.0561	0.0531
ARFIMA	0.0203	0.0154	0.4220	0.0675	0.0605	0.0198	0.0148	0.2040	0.0678	0.0569

The forecast evaluation measures calculated after fitting the scalar response model to the EUR-USD implied volatility data set. Delta=75, the in-sample period is January 2006 to December 2011, with the out-of-sample period spanning December 2011 to November 2013.

Table 4: Scalar response (SR), AR, GARCH, and ARFIMA models out-of-sample forecast evaluation measures (Delta=90)

Maturity			1 Month					3 Month		
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0068	0.0052	0.9460	0.0389	0.0320	0.0045	0.0034	0.9440	0.0311	0.0253
AR	0.0330	0.0256	0.5200	0.0885	0.0831	0.0255	0.0192	0.5020	0.0751	0.0692
GARCH	0.0352	0.0273	0.2960	0.0907	0.0879	0.0259	0.0195	0.4060	0.0739	0.0718
ARFIMA	0.0355	0.0276	0.2620	0.0936	0.0861	0.0265	0.0199	0.3200	0.0774	0.0698
Maturity			6 Month					9 Month		
ATM	RMSE	MAE	MCPDC	MME(O)	MME(U)	RMSE	MAE	MCPDC	MME(O)	MME(U)
SR	0.0041	0.0029	0.9380	0.0282	0.0233	0.0040	0.0027	0.9320	0.0269	0.0214
AR (1)	0.0223	0.0170	0.5000	0.0690	0.0668	0.0183	0.0139	0.5220	0.0607	0.0599
GARCH	0.0227	0.0172	0.4460	0.0679	0.0688	0.0188	0.0143	0.4420	0.0633	0.0594
ARFIMA	0.0237	0.0180	0.2340	0.0730	0.0672	0.0223	0.0170	0.1040	0.0725	0.0629

The forecast evaluation measures calculated after fitting the scalar response model to the EUR-USD implied volatility data set. Delta=90, the in-sample period is January 2006 to December 2011, with the out-of-sample period spanning December 2011 to November 2013.