Imperial College London
Imperial College Business School

Model Averaging for Volatility Forecasting, Option Pricing and Asset Allocation

Georgia Papadaki

February 2011

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Quantitative Finance of Imperial College London
and the Diploma of Imperial College London

Abstract

In this thesis the problem of model uncertainty is under scrutiny along with its implications in attaining optimal forecastability. To account for that averaging techniques are adopted including Bayesian model averaging, Bayesian Approximation and Thick Modelling. After an introductory chapter and a second one where some of the most celebrated conditional-volatility modelling proposals are discussed the third chapter investigates volatility forecasting and its direct association to option pricing. Some novel approaches to perform averaging are suggested here including variations of the predetermined methods together with more sophisticated algorithmic propositions such as Neural Networks. The fourth chapter extends the focal point of averaging to the whole predictive volatility density as this can be inferred first from derivatives on the underlying volatility index and second directly from the asset class under consideration (here the equity index) using bootstrap based - GARCH type models. The fifth chapter introduces some widely used variable selection techniques to the Finance continuum while averaging schemes once more are used in order to avoid model misspecification risk. Extensions to a nonlinear regression framework are also suggested while investment strategies are implemented in all chapters substantiating the ultimate supremacy of averaging schemes against single model alternatives. The last chapter concludes the research and makes some future suggestions for additional investigation.

Acknowledgments

I would like to express my appreciation and heartfelt gratitude to my supervisor Proffesor Paolo Zaffaroni for his inspiration, guidance and unconditional support throughout all these four years. I want to thank him for giving me this opportunity and reassure him that without him the completion of this thesis would not have been possible. I would also like to extend my acknowledgements to my second supervisor Professor Nigel Meade for his vital contribution and encouragement and for all the constructive discussions we had. His help in so much more ways than I can count is deeply appreciated. To both of them I am indebted for life.

I am also equally thankful to Professor William Perraudin who has made his support available in numerous instances as well as Simon Burbidge, High Performance Computer (HPC) coordination manager. Due to the complex and computationally intensive nature of the targeted applications his role has been a vital factor for the speed, accuracy and uniqueness of the results produced in the thesis. Regarding the use of HPC I owe friendly feelings of thankfulness to Dimosthenis Pediaditakis for his useful recommendations concerning any problems encountered. Also I am grateful to Professor Robert Bliss for all his comments and suggestions regarding the fourth chapter.

In addition I would like to make special references to my managers Vijay Chaudhary from BP plc and James Slaughter and Paul Hewett from Liberty Syndicates who have partially sponsored my research. I heartily feel obliged for their gracious professional collaboration and unequivocal understanding. Especially for the latter two I know a mere expression of how grateful I feel likewise does not suffice so I hope one day I will have the opportunity to reciprocate their generosity.

My sincere thanks are also due to the official examiners, Professor Stephen Taylor and Professor Nicholas Bingahm, for their detailed review, constructive criticism and excellent advice regarding this thesis. Their essential

assistance in revising this study has been of great value for me.

Last but by no means least I would like to take this moment to thank Fotis for his companionship and great patience at all times. I feel blessed that we walked together side by side during this process.

to my parents

Contents

1. Introduction **20**
1.1. Outline of the Thesis 21
1.2. Contributions 23

2. Conditional-Volatility Estimation & Forecasting **25**
2.1. Introduction 25
2.2. ARCH-type Volatility Models 25
2.2.1. Univariate Volatility Models 25
2.2.1.1. Autoregressive Conditional Heteroscedasticity process (ARCH) 25
2.2.1.2. Generalized Autoregressive Conditional Heteroscedasticity process (GARCH) 27
2.2.1.3. Leverage effect 29
2.2.1.4. GJR-GARCH & Threshold-GARCH process 30
2.2.1.5. Exponential-GARCH process (EGARCH) . . 30
2.2.1.6. ARCH-in-mean models 31
2.2.1.7. Intergraded GARCH (IGARCH) 31
2.2.1.8. Asymmetric Power ARCH model (APARCH) 32
2.2.1.9. Non-Linear ARCH process (NARCH) 33
2.2.1.10. Quadratic ARCH process (QARCH) 34
2.2.1.11. Component GARCH process (CGARCH) . . 35
2.2.2. Multivariate Volatility Models 36
2.2.2.1. The VECH model 36
2.2.2.2. The BEKK model 37
2.2.2.3. Factor Models 38
2.2.2.4. The Constant Conditional Correlation Model 40
2.2.2.5. The Dynamic Conditional Correlation Model 41
2.2.2.6. The Exponential Smother 43
2.2.2.7. Orthogonal GARCH 43

2.2.2.8. Generalized Orthogonal GARCH 44
2.2.2.9. Asymmetric Multivariate GARCH Models . 46
2.2.2.10. General Dynamic Covariance Matrix model 46
2.2.2.11. Asymmetric Dynamic Covariance Matrix model 47
2.2.2.12. Asymmetric VECH 48
2.2.2.13. Asymmetric CCC 48
2.2.2.14. Asymmetric BEKK 48
2.2.2.15. Asymmetric FARCH 48
2.2.2.16. Asymmetric Dynamic Conditional Correlation 49
2.2.2.17. Matrix Exponential GARCH 49
2.2.2.18. Modelling Multivariate Volatility with Copula-GARCH Models 50
2.2.2.19. Long Memory Volatility Models 50
2.2.2.20. Fractionally Integrated GARCH (FIGARCH) 51
2.2.2.21. Fractionally Intergraded EGARCH (FIEGARCH) 53
2.3. Stochastic Volatility Models 54
2.3.1. Univariate Stochastic Volatility Models 54
2.3.1.1. Mean-Reverting Stochastic Volatility models 55
2.3.1.2. The Vasicek Model 56
2.3.1.3. The Cox-Ingersoll-Ross (CIR) model 56
2.3.1.4. Heston's model 56
2.3.2. Multivariate Stochastic-Volatility Models 57
2.3.2.1. Introducing the fundamental model 58
2.4. Volatility-Forecasting Using High-Frequency Data and Implied Volatility 58
2.5. Estimation Methods 61
2.5.1. Markov Chain Monte Carlo 61
2.5.2. Bayesian Inference 63

3. Volatility Model Averaging and Option Pricing 68
3.1. Introduction 68
3.2. Literature Review 70
3.2.1. Bayesian Model estimation & averaging for conditional-volatility models 70
3.2.2. Thick Model Averaging for conditional-volatility models 71
3.2.3. Neural Networks and conditional-volatility forecasting 73

3.3. Methodology 75
3.3.1. Methodology for Bayesian Volatility Estimation 75
3.3.2. Methodology for Bayesian Model Averaging 79
3.3.3. Methodology for Bayesian Approximation for Model Averaging 81
3.3.4. Methodology for Thick Model Averaging 83
3.4. Volatility Model Averaging and Option Pricing 86
3.4.1. Literature Review 86
3.4.2. Methodology for Option pricing 88
3.4.3. Methodology for Averaging Option Prices against Averaging the underlying Volatility 91
3.5. Data 92
3.6. Empirical results 95
3.6.1. Empirical results for Volatility Forecasting 95
3.6.2. Empirical results for Option Pricing 98
3.6.2.1. Empirical results when Comparing with a Benchmark 101
3.6.2.2. Empirical results according to Maturity and Moneyness 101
3.7. Conclusions 103

4. Volatility Density Forecasting and Model Averaging 132
4.1. Introduction 132
4.2. Literature Review 137
4.2.1. Risk-Neutral Density Forecasting 137
4.2.2. Stochastic-Volatility Option Pricing 143
4.2.3. Transforming Risk-Neutral to Real-World Densities . . 154
4.3. Real-World Volatility Density Forecasting 157
4.3.1. Literature Review 157
4.4. Density Forecasting Evaluation 159
4.5. Methodology 163
4.5.1. Real-World Volatility-Density Forecasting 164
4.5.2. Risk-neutral Volatility-Density Forecasting (Option-based approach) 169
4.6. Density model averaging 174
4.6.1. Risk-Neutral Volatility-Density Averaging 175

4.6.2. Real-World Volatility-Density Averaging 178
4.7. Data 178
4.8. Results 180
4.8.1. Results of Risk-Neutral Volatility-Density Forecasting 180
4.8.2. Results: Real-World Volatility-Density Forecasting . . 182
4.8.3. Investment Strategies 183
4.9. Conclusions 184

5. Variable selection and application in Asset allocation problems 205
5.1. Introduction 205
5.2. Literature Review of Variable Selection Algorithms 209
5.3. Theoretical Framework 213
5.3.1. The Linear Variable Selection Algorithms 213
5.3.2. The Generalized Linear Variable Selection Algorithms 223
5.3.3. The Logistic Regression 224
5.3.4. ℓ_1 Regularized Logistic Regression 228
5.3.5. ℓ_2 Regularized Logistic Regression 230
5.3.6. Elastic Net Regularized Logistic Regression 231
5.3.7. SPCA and the Logistic Function 231
5.3.8. Model Averaging using Variable Selection Algorithms 232
5.4. Methodology 233
5.4.1. Accounting for a twofold uncertainty: *Cutting Point* on the Path and *Model* Uncertainty . . 239
5.5. Data 241
5.6. Results 244
5.6.1. Linear Results 244
5.6.2. Logit Results 247
5.6.3. Investment Strategies: Mean-Variance Analysis 250
5.6.3.1. Results - short sales restriction 252
5.6.3.2. Results - no short sales restriction 253
5.6.4. Logit based Analysis 254
5.6.4.1. Results - short sales restriction 254
5.6.4.2. Results - no short sales restriction 258
5.7. Conclusions 260

6. Conclusions & Further Research **296**

Appendix:

A. Additional Tables Chapter 3 **299**

B. Additional Tables Chapter 5 **305**

Bibliography **308**

List of Tables

3.1. Model Key Indicators and acronyms for Volatility tables. . . 105
3.2. Rankings according to RMSE of averaging methods and only those volatility models selected according to the best in-sample BIC criterion. 106
3.3. Table of the in-sample best performing models according to BIC criterion.All the GARCH-type models are examined up to lag 4. The best performing according to BIC information criterion is the one with the lowest BIC value. The last column gives in the squared brackets the (p,q) lag of this best model. 108
3.4. Table of the p-values generated by the Diebold-Mariano test. 110
3.5. Table including Diebold-Mariano comparison results. 112
3.6. Model Key Indicators and acronyms for Option tables. 114
3.7. Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods. 115
3.8. Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity less than 60 and Moneyness less than 1.05. 116
3.9. Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity less than 60 and Moneyness from 1.05 to 1.15. 117

3.10. Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity from 60 to 160 and Moneyness less than 1.05. 118
3.11. Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity from 60 to 160 days and Moneyness from 1.05 to 1.15. 119
3.12. Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity from 60 to 160 days and Moneyness more than 1.15. 120
3.13. Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity more than 160 days and Moneyness less than 1.05. 121
3.14. Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity more than 160 days and Moneyness from 1.05 to 1.15. 122
3.15. Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity more than 160 days and Moneyness more than 1.15. 123
3.16. Option prices descriptive statistics. 124
3.17. Table of the MLE and Bayesian based estimated model parameters (an indicative example). 124
3.18. Number of options on the S&P500 index used to calculate each surface. 125

4.1. Assumptions of the volatility process according to Detemple and Osakwe (2000). 145
4.2. The "work horses" of discrete-time conditional-volatility forecasting along with their restrictions. 158
4.3. Table of the GARCH-type model alternatives used for the estimation of the real-world volatility-forecasting density. . . 165
4.4. Table of the error distributions used in the analysis. 166
4.5. Rankings of the GARCH-type volatility density forecasting models according to the Empirical, Log-Normal and Gamma distribution. 185
4.6. Rankings of the risk-neutral volatility density forecasting models and averaging schemes. 186
4.7. Rankings of Objective volatility density forecasting models and averaging schemes. 187
4.8. Cumulative wealth generated by the investment strategy based on risk-neutral volatility density forecasting models and averaging schemes. 188
4.9. Cumulative wealth generated by the investment strategy based on Objective volatility density forecasting models and averaging schemes. 189
4.10. Rankings of all risk-neutral volatility density forecasting models and averaging schemes. 191
4.11. Rankings of all Objective volatility density forecasting models and averaging schemes. 193
4.12. Rankings of the first 60 out of 334 GARCH-type real-world volatility density forecasting models according to the empirical distribution. 194
4.13. Rankings of the first 60 out of 334 GARCH-type real-world volatility density forecasting models according to the Log-Normal distribution. 196
4.14. Rankings of the first 60 out of 334 GARCH-type real-world volatility density forecasting models according to the Gamma distribution. 198
4.15. Cumulative wealth generated by the investment strategy based on all risk-neutral volatility density forecasting models and averaging schemes. 199

4.16. Cumulative wealth generated by the investment strategy based on all Objective volatility density forecasting models and averaging schemes. 201
4.17. Number of options on the VIX index used to calculate each surface. 202

5.1. Rankings of the first 10 best Linear Models and averaging schemes for all asset classes. 262
5.2. Rankings of the first 10 best Logit Models and averaging schemes for all asset classes. 263
5.3. Rankings of of the first 10 best Linear Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction. 264
5.4. Rankings of the first 10 best Linear Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction. . . . 265
5.5. Rankings of asset allocation strategies for EQUITIES of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction. 266
5.6. Rankings of asset allocation strategies for BONDS of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction. 267
5.7. Rankings of asset allocation strategies for COMMODITIES of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction. 268
5.8. Rankings of asset allocation strategies for FOREIGN EXCHANGE of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction. 269
5.9. Rankings of asset allocation strategies for EQUITIES of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction. 270

5.10. Rankings of asset allocation strategies for BONDS of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction. 271
5.11. Rankings of asset allocation strategies for COMMODITIES of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction. 272
5.12. Rankings of asset allocation strategies for FOREIGN EXCHANGE of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction. . . . 273
5.13. Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restrictions for EQUITIES. 275
5.14. : Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restrictions for BONDS. 276
5.15. : Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restrictions for COMMODITIES. . . 278
5.16. : Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restrictions for FOREIGN EXCHANGE. 279

5.17. Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restrictions for EQUITIES. 280
5.18. Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restrictions for BONDS. 282
5.19. Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restrictions for COMMODITIES. 283
5.20. Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restrictions for FOREIGN EXCHANGE. 285
5.21. Diebold-Mariano test for EQUITIES at 95% confidence level. 286
5.22. Diebold-Mariano test for BONDS at 95% confidence level. . . 288
5.23. Diebold-Mariano test for COMMODITIES at 95% confidence level. 290
5.24. Diebold-Mariano test for FOREIGN EXCHANGE at 95% confidence level. 292

A.1. Summary statistics of S&P 500 for the period 31_{st} August 1997 to 1_{st} September 2007. 299
A.2. Rankings of all volatility models and averaging methods according to RMSE. 302

A.3. Rankings according to $RMSE of all option pricing models and Model Averaging methods. 304

B.1. Description and details regarding the data used in chapter 5. 307

List of Figures

3.1. Plot of the fitted distribution of GARCH(1,2) parameters. This is the actual distribution of each model parameter of the GARCH(1,2) along with the Normal fitted proposal distribution. The ARCH and the first GARCH parameter clearly resemble the proposal Normal . The same however does not hold here for the rest of the parameters. 126
3.2. Convergence Plots of parameters to their theoretical distribution using MCMC. This is the process of convergence for all parameters of NGARCH(3,4), NAGARCH(2,2) and GJR-GARCH(2,2) to their actual distribution using MCMC with every line standing for a different parameter. 127
3.3. Convergence Plots of parameters to their theoretical distribution using MCMC. This is the process of convergence for all parameters of GARCH(3,2), AVGARCH(4,2), APGARCH(1,1).128
3.4. Plot of two proxies, squared returns and high-low estimator and GARCH(1,1) for the out-of-sample period (260 days or else one year from 31/8/2006 to 1/9/2007). 129
3.5. Plot of the out-of-sample performance of Duan's option pricing model against and Thick Modelling with weights based on option $RMSE and 20% retention percentage. 130
3.6. Implied volatility surface based on options on the S&P500 index. 131

4.1. Plot of the cumulative wealth generated by the investment strategy based on Objective volatility density forecasting processes under all three distributional assumptions and in particular for the Best single models selected according to BIC, for the Thick Model Averaging method using BIC and for the S&P500 index. 203

4.2. Implied volatility surface based on options on the VIX index. 204

5.1. LARS shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the linear regression application on equities. Each line represents a coefficient. 293
5.2. LASSO shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the linear regression application on equities. . . 293
5.3. Elastic Net shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the linear regression application on equities. 294
5.4. LASSO shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the logit regression application on equities. . . . 294
5.5. Ridge regression shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the logit regression application on equities. 295
5.6. Elastic Net shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the logit regression application on equities. 295

1. Introduction

It is widely accepted in the modern finance theory that out-of-sample forecasting performance is the ultimate objective of any model aspiring to substantiate its supremacy among other competing alternatives. From a simple autoregressive process for time series to artificial intelligence algorithms, dispute still divides the scientific community regarding which method can yield the best result. Amid these controversies volatility modelling is of primary importance not only because of its extensive interrelationship with the vast majority of financial products but also because of its unique characteristics. In contrast with other indicators whose value is directly observable, volatility is a latent variable. This in turn implies that unbiased proxies have to be used instead, which makes model validation difficult.

Since the second central moment has been modeled extensively there is a plethora of choices all trying to capture particular irregularities connected to conditional-volatility forecasting such as clustering and leverage effects. Identifying in-sample the candidate scheme that yields the best predictive performance is rarely possible due to model misspecification risk. Additionally the decision environment is far from stable due to the complexity of capital markets, the continuous introduction of new financial products and the enormous amount of financial data generated on a daily basis. Heterogeneities are present and what is optimal today may not be tomorrow.

In addition advances in contemporary risk management suggest progression from moments of the distribution to the whole predictive density. The recent introduction of derivatives on some volatility indexes offers a further incentive towards that goal. In order to move beyond model uncertainty, irrespective of our focus, averaging techniques can offer some help. We propose in this thesis to try to eliminate any biases by averaging all, the majority or just the best-performing model alternatives instead of selecting a single candidate to perform forecasting. Although the idea dates back to the 1960s, it has yet to receive proper recognition of its importance due

to unavoidable computational intensity involved in the implementation process. But contemporary technological advances make averaging methods far more applicable.

The main purpose of this research is to provide a better understanding of the benefits of averaging. In addition we hope to lay the foundations to encourage future scientific development of the method to benefit other financial applications.

As well as model selection, we also consider variable selection. Although the initial focus of the thesis is volatility forecasting, the primary objective is improving predictability; thus we also concentrate on the variables-indexes themselves. Model and variable ambivalence is addressed again with averaging related techniques, all in an attempt to offer superior forecasting results.

1.1. Outline of the Thesis

The thesis consists of six chapters: two introductory ones, three that comprise the body of the research which tries to improve forecasting by removing the barriers created by model uncertainty, and a final chapter that recapitulates the findings and suggests future improvements. Initially we focus on the effect of model averaging on option pricing, and afterwards on asset allocation and risk management related issues. Finally our attention turns to variable selection and predictability. In more detail, the thesis is structured as follows.

At first the most well established models used for time series forecasting, with a particular emphasis on conditional-volatility, are introduced. Starting with ARCH-type discrete-time models, we also consider stochastic volatility alternatives. The chapter concludes with some further details regarding the estimation process. The majority of the models presented in this part are used extensively in all subsequent sections. This review is conducted from a univariate as well as a multivariate perspective.

The third chapter addresses the problem of acquiring accurate volatility forecasts amidst an overabundance of competing models. This is also directly related to option pricing. Here some of the most popular averaging methods, including Bayesian Model averaging, Bayesian Approximation and Thick Modelling, are used to account for the problem of model se-

lection. Supplementary extensions of these classical averaging approaches are introduced. Thick Model Averaging using an optimisation algorithm to determine the loadings of every model as well as Thick Model Averaging using Neural Networks are proposed in an attempt to capture any non-linear relationships when combining the competing propositions. An empirical analysis based on the S&P 500 index is performed to test whether these methods can provide more accurate volatility forecasts compared to individual models. The different averaging volatility schemes are afterwards used for pricing options on the same index, addressing an important gap in the literature; the assumption that a particular volatility model drives the dynamics of the underlying asset. In the context of discrete volatility option pricing, two further innovations are suggested and contrasted. The first introduces pricing by averaging the individual option prices under different volatility dynamics. The second suggestion refers to a pricing method where first the alternative volatilities processes are averaged and afterwards this averaged volatility scheme alone is used for pricing. Both proposals are compared against single alternatives. The empirical analysis indicates a greater accuracy of the averaging methods when forecasting conditional-volatility as well as when we price an option.

In the fourth chapter, in addition to forecasting the second moment, the predictive density as a whole becomes the centre of the research. There the different volatility density forecasting methods are under investigation along with model averaging. The focus first is on real-world densities using bootstrap-based GARCH-type models on the same index (S&P 500). Additionally we are interested in the risk-neutral predictive density of volatility as this is derived from option prices on the VIX index instead. The total number of alternative option-pricing models used for density forecasting purposes is large, highlighting this way the difficulties created by model uncertainty. Empirically we demonstrate that there is no distinctive optimal choice that can capture all the attributes of the underlying asset's distribution. By combining all or just the top performing models together one can achieve even higher forecastability. Here Thick Modelling and Bayesian Approximation are used in order to perform averaging. Additionally we implement a volatility trading strategy based on the forecasted densities and we show that, both econometrically and empirically, averaging helps significantly to increase performance and ultimately profitability.

The fifth chapter increases the difficulty of identifying the best predictive model by adding the uncertainty of variable selection. Here notable algorithms are used to facilitate the process including Ridge regression, all-subsets regression, LASSO, LARS, Elastic Net and Sparse Principal Components Analysis, to mention just a few. The empirical analysis is centered around nineteen candidate predictors, among which we try to identify and select only those that have the most economic significance. These are used afterwards as a forecasting vehicle in order to achieve predictive supremacy. Additionally, model selection is addressed again by averaging methods, including as before Bayesian Approximation and Thick Modelling. A further source of uncertainty had also been taken under consideration, which relates to the way the advanced algorithmic techniques used for variable selection are structured. In more detail, each individual algorithmic procedure when executed generates paths so that at each step a different model choice is available. The researcher must identify the optimal cutting point on these paths which represents the best single model to be used for his analysis. This adds an additional uncertainty to the variable selection process. To account for all the above we have first to perform averaging to the individual steps on the paths so that cutting point bias is eliminated and second to average across all paths that all models produce so that model uncertainty is addressed as well.

The chapter concludes with empirical applications which extend the scope of this part to a more general logistic regression framework. Instead of predicting asset returns, the investor's only interest might be to forecast the direction of the movement. The investment strategies employed here utilize the outcome of both the linear and the logistic regression. Results indicate once more dominance of the averaging schemes over any single model alternative.

The last chapter summarizes the findings of the research while some final conclusions are provided as well. The thesis ends with various indications for future investigation.

1.2. Contributions

Our analysis puts into a whole new perspective the concept of averaging and emphasizes its significant advantage over single model selection. The imple-

mentation of the method is based mainly on new computational advances from which the researcher can now reap the benefits. As we demonstrate, model uncertainty impacts the decision making process at all levels, with potential hazardous effects on profitability. By adopting an averaged model scheme we provide empirical evidence that the investor can not only be more assertive regarding his portfolio structure, but also increase his potential of attaining investment targets. A further advantage of the method is its versatility, since there exists an abundance of financial applications where averaging can be adopted. Demonstrating this potential, some novel applications were executed, including option pricing, density forecasting and algorithmic variable selection. All these unavoidably lead to more optimal strategies for asset allocation. Along with the introduction of some alternative methods for averaging, the uniqueness of this research lies also in the scale of the various problems involved. This relates not only to the number of models considered and the size of the investment horizon employed, but most importantly to the variety of asset classes used for forecasting purposes. Finally, we seek to increase the awareness of the averaging alternatives that exist, and what they have to offer towards improving the prediction outcome. We highlight their advantages and disadvantages in various financial applications, and ultimately which one of them exhibits the most consistent performance.

Finally chapter five also introduces the ability of Commodity and Foreign Exchange indexes to predict not only related asset classes but also Stock and Bond prices. However, further investigation is suggested, although empirical evidence is in favour of this claim.

2. Conditional-Volatility Estimation & Forecasting

2.1. Introduction

In this chapter some of the most popular models used for conditional-volatility forecasting are presented both from a univariate and a multivariate perspective. We treat separately discrete-time and continuous-time alternatives. All are scrutinized, emphasizing model-specific properties which account for some of the stylized attributes accompanying volatility. Reference is also made to the various estimation methods that exist and are used in order to facilitate in the parameter estimation process. The chapter concludes with some alternative suggestions to accurately capture future movements of the second distributional moment including model averaging which is the center of all the thesis.

2.2. ARCH-type Volatility Models

2.2.1. Univariate Volatility Models

2.2.1.1. Autoregressive Conditional Heteroscedasticity process (ARCH)

Let a time series process y_t be defined as an ordered set of realizations of some random variable sorted according to time. An Autoregressive process of order p or AR(p) is one for which the value of a time series today equals that of the previous day (a portion of it in particular) in addition to some constant and an innovation term.

$$y_t = \varphi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t \quad \varepsilon_t \sim WN(0) \tag{2.1}$$

where WN stands for White Noise and p is a non negative integer such that the past p values of a time series y_{t-i} $\forall i = 1, \ldots, p$ jointly determine the future expectation of y_t. Properties of y_t are completely defined by ϕ and the variance of the innovation process h_t. Additionally let us define ε_t as a shock progression that can be model as an AR process.

$$\varepsilon_t^2 = \omega + a\varepsilon_{t-1}^2 + \eta_t \quad \eta_t \sim WN(0) \tag{2.2}$$

The variance of the process ε_t then equals

$$\begin{gathered} V[\varepsilon_t] = E_t[\varepsilon_t^2] \equiv h_t, \\ E_t[\omega + a\varepsilon_{t-1}^2 + \eta_t] = \omega + a\varepsilon_{t-1}^2, \\ h_t = \omega + a\varepsilon_{t-1}^2. \end{gathered} \tag{2.3}$$

More generally,

$$h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2. \tag{2.4}$$

This is the benchmark of every subsequent attempt to model volatility and it has been proposed by Engle in 1982. To ensure that the conditional variance is strictly positive the parameter values should be restricted as follows:

$$\begin{aligned} &\omega > 0, \\ &a_i \geq 0 \quad \forall i = 1, \ldots, p. \end{aligned} \tag{2.5}$$

Additionally the process is covariance stationary if and only if the roots of the polynomial $1 - a_1 L - \ldots - a_p L^p = 0$ lie outside of the unit circle. Thus,

$$\sum_{i=1}^{p} a_i < 1. \tag{2.6}$$

The unconditional variance of y_t equals,

$$\sigma^2 = \frac{\omega}{1 - \sum\limits_{i=1}^{p} a_i}. \tag{2.7}$$

Substituting in h_t, we get

$$h_t = \sigma^2(1 - \sum_{i=1}^{p} a_i) + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2. \tag{2.8}$$

Finally, for the error term,

$$\begin{gathered} V[\varepsilon_t] = E[\varepsilon_t^2] = E[E_{t-1}[\varepsilon_t^2]] = E[h_t], \\ V[\varepsilon_t] = \frac{\omega}{1 - \sum_{i=1}^{p} a_i}. \end{gathered} \tag{2.9}$$

2.2.1.2. Generalized Autoregressive Conditional Heteroscedasticity process (GARCH)

Another widely used process to forecast time series is the Moving Average model of order q. This simply implies that next day's value equals a constant plus a weighted average of shocks $\varepsilon_{t-i}\ \forall i = 1, \ldots, p$. In general,

$$y_t = \varphi_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \quad \varepsilon_t \sim WN(0). \tag{2.10}$$

MA models are always weakly stationary because they are linear combinations of a white noise process, of which the first two moments are time invariant. The estimation of high-order models definitely requires the parameters to be adequately described by the dynamics of their structure. Researchers found it extremely burdensome to use AR and MA models separately thus as a natural extension they combined the two yielding the ARMA model of order (p,q):

$$\begin{gathered} y_t = \varphi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}, \\ \varepsilon_t \sim WN(0). \end{gathered} \tag{2.11}$$

Properties of an ARMA (p,q) model include the weak stationary condition which is satisfied if all the solutions of the characteristic equation are less than 1 in absolute value. If we now assume that shocks follow an ARMA process, then

$$\varepsilon_t^2 = \omega + a\varepsilon_{t-1}^2 + \eta_t + \beta\eta_{t-1}, \quad \eta_t \sim WN(0), \tag{2.12}$$

$$
\begin{aligned}
V[\varepsilon_t] &= E_t[\varepsilon_t^2] \equiv \sigma_t^2, \\
E_t[\omega + \gamma\varepsilon_{t-1}^2 + \eta_t + \delta\eta_{t-1}] &= \omega + \gamma\varepsilon_{t-1}^2 + \delta\eta_{t-1} \\
&= \omega + \gamma\varepsilon_{t-1}^2 + \delta(\varepsilon_{t-1}^2 - E_{t-2}[\varepsilon_{t-1}^2]) \\
&= \omega + \gamma\varepsilon_{t-1}^2 + \delta(\varepsilon_{t-1}^2 - \sigma_{t-1}^2) \\
&= \omega + (\gamma + \delta)\varepsilon_{t-1}^2 - \delta\sigma_{t-1}^2. \qquad (2.13)
\end{aligned}
$$

Substituting $\gamma + \delta = a$ and $-\delta = \beta$ we can get a more general form of the equation:

$$
h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}. \qquad (2.14)
$$

The GARCH model has been introduced by Bollerslev in 1986 and is reasonably characterized as the workhorse of financial time series analysis. However, a number of properties have to be addressed in order to assure positive variance and covariance stationarity.

- $\omega > 0,\ a_i, \beta_j \geq 0 \quad \forall i = 1, \ldots, p, \quad j = 1, \ldots, q,$
- $\beta_j = 0$ if $a_i = 0$,
- $\sum\limits_{i=1}^{p} a_i + \sum\limits_{j=1}^{q} \beta_j < 1.$

The unconditional variance then equals

$$
V[\varepsilon_t] = E[h_t] = \frac{\omega}{1 - \sum\limits_{i=1}^{p} a_i - \sum\limits_{j=1}^{q} \beta_j}. \qquad (2.15)
$$

In ARCH and GARCH models the parameters are usually calculated using the Maximum Likelihood Estimation (MLE), while the assumption of a normal distribution for the residuals is also commonplace. However, different distributional patterns can also be employed. Let us assume the following:

$$
\varepsilon_t \,|F_t \sim N(0, h_t)\,, \qquad (2.16)
$$

where $\theta \equiv [\omega,\ \alpha_i,\ \beta_j]\ \forall i = 1, \ldots, p$ and $\forall j = 1, \ldots, q$ are the unknown parameters to be estimated via MLE. Then

$$\begin{aligned} L(\theta\,|y_1, y_2, \ldots, y_t) &= \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi h_t}} \exp(-\frac{\varepsilon_t^2}{2h_t}) \\ &= -\frac{1}{2}\{(T-1)\log(2\pi) + \sum_{t=2}^{T} \log(h_t) + \sum_{t=2}^{T} \frac{\varepsilon_t^2}{h_t}\}. \end{aligned} \tag{2.17}$$

If we want to make no assumption about the distributional pattern of the residuals then a Quasi Maximum Likelihood Estimator can be used instead (QMLE), where estimates of the parameters are obtained using numerical optimisation. Bollerslev and Wooldridge (1992) argue however that "using the normal distribution in maximum likelihood will give us consistent parameter estimates even if the true density is non-normal". Thus an assumption of normality, even if it is wrong, can yield trustworthy parameter estimates when using ML; however, QML is a consistent and asymptotically normal estimator.

Intrinsic weaknesses of the previous two volatility models (ARCH, GARCH), such as equal treatment of positive and negative shocks, inability to capture excess kurtosis, insufficient understanding of the source of variation and slow response to large isolated shocks, laid the groundwork for a variety of extensions that followed.

2.2.1.3. Leverage effect

It has long been an article of faith (see Black (1976)) that volatility responds differently to the announcement of bad and good news. More precisely, volatility has the tendency to rise more following negative shocks than positive shocks. The News Impact Function, also known as the News Impact Curve (Engle and Ng (1993)), has been proposed in the past as a tool to measure volatility irregularity. A series of models also made their appearance in an attempt to deal with this abnormality. The most widely used will be discussed here.

2.2.1.4. GJR-GARCH & Threshold-GARCH process

Glosten, Jagannathan and Runkle (1993) extend the traditional GARCH model to accommodate the need to account for a possible asymmetric impact on conditional volatility of positive and negative shocks. What they propose is the GJR-GARCH model, which can be expressed as follows,

$$h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \gamma \varepsilon_{t-i} I\{\varepsilon_t < 0\} + \sum_{j=1}^{q} \beta_j h_{t-j}, \tag{2.18}$$

where the variable $I(x)$ is defined as

$$I(x) = \begin{cases} 1 & x \leq 0, \\ 0 & x > 0. \end{cases} \tag{2.19}$$

To ensure positive conditional variance we restrict parameters as follows,

- $\omega > 0$,
- $a_i > 0 \; \forall i = 1, \ldots, p$,
- $a_i + \gamma > 0 \; \forall i = 1, \ldots, p$,
- $\beta_j \geq 0 \; \forall j = 1, \ldots, q$.

If $\gamma > 0$ the impact of today's return on tomorrow's volatility is greater if the return is negative.

The only difference between Threshold GARCH (Zakoian (1994)) and GJR is that it captures the standard deviation instead of the variance while the model uses absolute residuals instead of squared.

$$h_t^{1/2} = \omega + \sum_{i=1}^{p} a_i \left|\varepsilon_{t-i}\right| + \gamma \left|\varepsilon_{t-i}\right| I\{\varepsilon_{t-i} < 0\} + \sum_{j=1}^{q} \beta_j h_{t-j}^{1/2}. \tag{2.20}$$

2.2.1.5. Exponential-GARCH process (EGARCH)

Nelson (1991) suggests a model that alleviates the restrictive nature of GARCH, first by tackling the conditions imposed on the parameters to ensure positive variance, and second by accounting for leverage effects. His proposition captures variance as below.

$$\log(h_t) = \omega + \sum_{i=1}^{p} a_i \left| \frac{\varepsilon_{t-i}^2}{h_{t-i}} \right| + \sum_{k=1}^{r} \gamma_k \frac{\varepsilon_{t-k}}{h_{t-k}} + \sum_{j=1}^{q} \beta_j \log(h_{t-j}). \tag{2.21}$$

Unlike its predecessors, for EGARCH variance lies in the real line ($\log(h_t)$) and any restriction imposed on the parameters to guarantee positive definiteness is redundant. More importantly EGARCH can effectively account for numerous stylized facts of volatility such as asymmetric implications of different signed shocks. Additionally EGARCH uses standardized rather than unconditional shocks which have finite moments. However, as EGARCH model evolves in a nonlinear way for higher-order models, nonlinearity becomes too complicated (Tsay (1992)).

2.2.1.6. ARCH-in-mean models

In 1987 Engle, Lilien and Robins suggested a model that could clearly demonstrate the property of financial returns to positively correlate with risk. Thus the conditional mean of returns here is also a function of its conditional variance. The dynamics that govern the conditional variance are subjective and depend on the individual preferences of the researcher. It is suggested however to use the standard deviation instead of the variance when modelling since the former is measured in the same units as returns. For example the following representation could belong to this family:

$$r_t = \phi_0 + \phi_1 r_{t-1} + \delta h_t^{1/2} + \varepsilon_t, \tag{2.22}$$

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1}. \tag{2.23}$$

The parameter δ is called the risk premium, and thus a positive δ demonstrates a positive correlation between risk and return. It also implies the existence of serial correlation between past returns which is introduced by the volatility.

2.2.1.7. Intergraded GARCH (IGARCH)

A fundamental property of GARCH models is that they are mean-reverting. There are cases, however, when mean reversion is slow or may not happen

at all, and this can be easily achieved by imposing $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} \beta_j = 1$. This in turn implies that the volatility is non-stationary or has a unit root. In other words, IGARCH models are GARCH models with unit root. What characterizes IGARCH models is the persistence of past squared shocks. The model can be written as below, and it was first introduced by Engle and Bollerslev in 1986.

$$h_t = \omega + \sum_{j=1}^{q} (1 - \beta_j)\varepsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}. \tag{2.24}$$

2.2.1.8. Asymmetric Power ARCH model (APARCH)

Ding, Granger and Engle (1993) offer a generalization to Taylor's model (1986) which related the future conditional-volatility to past absolute residuals and standard deviations, and propose not an imposition of a structure on data but on the "optimal power term". Consequently the model offers a wide range of transformation capabilities including ARCH and GARCH models for conditional volatility estimates.

$$\sqrt[\gamma]{h_t} = \omega + \sum_{i=1}^{p} a_i \left(|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i}\right)^{\gamma} + \sum_{j=1}^{q} \beta_j \sqrt[\gamma]{h_{t-j}}. \tag{2.25}$$

Some of the most widely known extensions of APARCH can be summarized in the following seven categories (Laurent, 2003).

1. ARCH, Engle (1982) when $\gamma = 2,\ \gamma_i = 0\ \forall i = 1, \ldots, p,\ \beta_j = 0\ \forall j = 1, \ldots, q$.
2. GARCH, Bollerslev (1986) when $\gamma = 2,\ \gamma_i = 0\ \forall i = 1, \ldots, p$.
3. Taylor (1986)/Schwert (1990) GARCH when $\gamma = 1,\ \gamma_i = 0\ \forall i = 1, \ldots, p$.
4. GJR, Glosten, Jagannathan, and Runkle (1993) when $\gamma = 2$.
5. TARCH of Zakoian (1994) when $\gamma = 1$.
6. NARCH, Higgins and Bera (1992) when $\gamma_i = 0\ \forall i = 1, \ldots, p,\ \beta_j = 0\ \forall j = 1, \ldots, q$.
7. Log-ARCH, Geweke (1986) and Pentula (1986), when $\gamma \to 0$.

The "optimal power term" here is γ, which can be estimated rather than imposed, while γ_i is used to capture the impact of asymmetric shocks to volatility. The shocks that are used are unconditional rather than standardized (see EGARCH). The intuition behind the model can be traced in the fact that sometimes absolute returns may serve better than squared returns in our attempt to forecast volatility; thus there is no reason not to include them instead (or even any power representation of the returns).

2.2.1.9. Non-Linear ARCH process (NARCH)

If we begin with the general Asymmetric Power ARCH model (Ding *et al.*), the conditional-volatility equals

$$\sqrt[\gamma]{h_t} = \omega + \sum_{i=1}^{p} a_i \left(|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i}\right)^{\gamma} + \sum_{j=1}^{q} \beta_j \sqrt[\gamma]{h_{t-j}}. \tag{2.26}$$

Higgins and Bera (1992) set $\gamma,\ a_i$ parameters free and $\gamma_i,\ \beta_i = 0$. The model that they have designed is as follows:

$$h_t \quad = \quad (\phi_0 \omega^{\delta} + \phi_1 \varepsilon_{t-1}^{2\delta} + \ldots + \phi_p \varepsilon_{t-p}^{2\delta})^{1/\delta}, \tag{2.27}$$

where $\omega \geq 0,\ \phi_i \geq 0\ \forall i = 0, \ldots, p,\ \delta > 0$ and $\sum_{i=0}^{p} \phi_i = 1$.

As the authors point out the process has $p + 3$ parameters to estimate, while restricting ϕ_i to sum to one reduces the computation burden. Thus we end up with just one more parameter than ARCH to estimate. In addition, δ is of extreme importance, as restricting the parameter to be greater than zero ensures the actual existence of the conditional variance. It can also be used to express the elasticity of substitution, $\frac{1}{1-\delta}$, for $0 < \delta \leq 1$, which in turn can transform h_t equation to account for linearizations of traditional non-linear models as demonstrated below.

$$\frac{h_t^{\delta} - 1}{\delta} = \phi_0 \frac{\omega^{\delta} - 1}{\delta} + \phi_1 \frac{\varepsilon_t^{2\delta} - 1}{\delta} + \phi_2 \frac{\varepsilon_{t-1}^{2\delta} - 1}{\delta} + \ldots . + \phi_p \frac{\varepsilon_{t-p+1}^{2\delta} - 1}{\delta}. \tag{2.28}$$

This is a Box-Cox (1964) power transformation of Engle's ARCH equation.

2.2.1.10. Quadratic ARCH process (QARCH)

QARCH model is an attempt to combine many of the attractive features of the already existed ARCH models while ensuring non negativity of the conditional variance. Sentana (1991) defines the model as follows.

Let $\mu(X_{t-1\infty}) = 0$ and $h(X_{t-1\infty}) = 0$ be a function of the past q values of x_t only. This implies $h(X_{t-1\infty}) = h(X_{t-1q})$, and the estimated variance equals

$$\begin{aligned} h(X_{t-1q}) &= \theta + \sum_{i=1}^{q} \psi_i x_{t-i} + \sum_{i=1}^{q} a_{ii} x_{t-i}^2 + 2 \sum_{i=1}^{q} \sum_{j+1}^{q} a_{ij} x_{t-i} x_{t-j} \\ &= \theta + \psi' X_{t-1,q} + X'_{t-1,q} A X_{t-1,q}. \end{aligned} \tag{2.29}$$

ψ is a $q \times 1$ vector and A a symmetric $q \times q$ matrix. The above model entails the quadratic variance proposed in the Augment ARCH model (Bera and Lee (1990)) as an alternative to the ARCH parametric variance function, the linear standard deviation model (Robinson (1991)) and the Asymmetric ARCH (Engle (1990)). In the QARCH perspective ψ is allowed to vary and take any value rather than be confined to be centered around zero (in the case of ARCH and AARCH ψ=0). As a result the parameter can be used to adequately capture asymmetries and leverage effects. To derive the conditions necessary to ensure a positive-definite variance, we rewrite the model as follows:

$$h(X_{t-1q}) = (X'_{t-1q}, 1) \begin{bmatrix} A & \frac{\psi'}{2} \\ \frac{\psi'}{2} & A \end{bmatrix} \begin{bmatrix} X_{t-1q} \\ 1 \end{bmatrix}. \tag{2.30}$$

This is positive if the matrix $\begin{bmatrix} A & \frac{\psi'}{2} \\ \frac{\psi'}{2} & A \end{bmatrix}$ and thus A is positive semi definite.

An extension to the above model is the GQARCH which nests the GARCH (Bollerslev (1986)) and GAARCH (Bera and Lee (1990)) and is the following:

$$h_t = \frac{\theta}{1-\delta_1} + \psi_1 x_{t-1} + a_{11} x_{t-1}^2 + \sum_{j=2}^{\infty} (\psi_2 + \delta_1 \psi_1) \delta_1^{j-2} x_{t-j} + \\ + \sum_{j=2}^{\infty} (a_{22} + \delta_1 a_{11}) \delta_1^{j-2} x_{t-j}^2 + 2 \sum_{j=2}^{\infty} a_{12} \delta_1^{j-2} x_{t-j} x_{t-j-1}. \tag{2.31}$$

To ensure covariance stationarity the parameters a, δ should be restricted such that $\sum_{i=1}^{p} a_{ii} + \sum_{j=1}^{q} \delta_j < 1$ and its unconditional variance equals $V(x_t) = \frac{\theta}{\sum_{i=1}^{p} a_{ii} + \sum_{j=1}^{q} \delta_j}$.

2.2.1.11. Component GARCH process (CGARCH)

GARCH models provide limited freedom in their attempt to capture both the conditional and the unconditional distribution as parameters restrict the conditional variance to be governed by its unconditional value. An alternative to address for this inadequacy has been introduced by Ding and Granger (1996) and Engle and Lee (1999). A simple illustration is if you consider a two-component GARCH, CGARCH(2) (it can be easily extended additional components),

$$h_t = h_{t,1} + h_{t,2}, \tag{2.32}$$

$$h_{t,1} = \omega + a_1 \varepsilon_{t-1} + \beta_1 h_{t-1,1}, \quad \varepsilon_t \sim WN(0, \sigma^2), \tag{2.33}$$

$$h_{t,2} = a_2 \varepsilon_{t-1} + \beta_2 h_{t-1,2}. \tag{2.34}$$

For identification purposes we impose $\beta_1 > \beta_2$. To ensure positive conditional variance $1 > \beta_1 > \beta_2 > 0,\ \beta_2 > a_1 + a_2,\ \omega > 0,\ a_1 > 0,\ a_2 > 0$, and they are also sufficient conditions to assure covariance rationality. The difference in the above model can easily be traced to the fact that each volatility component reacts in a different way to the most recent innovation and has dissimilar rate of decay to lagged values. Thus one component can be used to capture the long-run effect of shocks and the second the short-run effect and vice versa. The effect of the past innovations can be depicted as follows:

$$h_{t,1} = \frac{\omega}{1-\beta_1} + a_1(\varepsilon_{t-1} + \beta_1\varepsilon_{t-2} + \beta_1^2\varepsilon_{t-3} + \ldots), \quad (2.35)$$

$$h_{t,2} = a_2(\varepsilon_{t-1} + \beta_1\varepsilon_{t-2} + \beta_1^2\varepsilon_{t-3} + \ldots), \quad (2.36)$$

$$\frac{\partial h_t}{\partial \varepsilon_{t-\kappa}} = a_1\beta_1^{k-1} + a_2\beta_2^{k-1}. \quad (2.37)$$

The unconditional variance equals

$$E(h_t) = E(h_{t,1}) + E(h_{t,2}) = \frac{\omega}{1-\beta_1}. \quad (2.38)$$

This in turn implies that the first component only determines the unconditional variance (the second equals 0), so it is considered as the long-run component. Even though CGARCH loses restrictions on parameters regarding the unconditional variance and employs a more flexible structure to capture the slowly decaying autocorrelation function, it still implies a restricted GARCH model. Thus the individual components cannot be independently identified.

2.2.2. Multivariate Volatility Models

2.2.2.1. The VECH model

Bollerslev *et al.* (1998) propose a model specifically designed to estimate multivariate volatility problems as ARMA type processes. To begin with let us consider a $\kappa \times 1$ vector of the stochastic process y_t, where $y_t = \mu_t(\theta) + \varepsilon_t$ with θ being the parameter vector and $\mu_t(\theta)$ the conditional mean vector. Also $\varepsilon_t = H_t^{1/2}(\theta)z_t$, with $H_t^{1/2}(\theta)$ being a $\kappa \times \kappa$ positive definite matrix where H_t is the conditional variance matrix and z_t is a random vector with its first two moments defined as $E(z_t) = 0$, $Var(z_t) = I_\kappa$ (I_κ is the identity matrix of order κ). The VECH model is formulated as follows:

$$vech(H_t) = C + \sum_{i=1}^{p} A_i vech(\varepsilon_{t-i}\varepsilon'_{t-i}) + \sum_{j=1}^{q} B_j vech(H_{t-j}) \quad \varepsilon_t \,|\psi_{t-1} \sim N(0, H_t). \quad (2.39)$$

The vech operator takes any symmetric matrix ($\kappa \times \kappa$) and returns its lower diagonal elements placed into a $\frac{\kappa(1+\kappa)}{2}$ vector. The matrices A, B are

of dimension $\frac{\kappa(1+\kappa)}{2} \times \frac{\kappa(1+\kappa)}{2}$ while C is of dimension $\frac{\kappa(1+\kappa)}{2} \times 1$. As a result the total number of parameters to be estimated is of order κ^4 or $O(\kappa^4)$. This implies that the the coefficients grow exponentially as the sample increases at a linear rate. Thus we have to keep κ very small while further restrictions have to be imposed on the parameter matrices A, B. Bollerslev *et al.* (1998) has also suggested the Diagonal VECH (DVECH), where A, B are diagonal matrices, which reduces the estimation burden to $\frac{\kappa(\kappa+5)}{2}$ parameters. But the model still entails a considerable amount calculations to be performed, and as a result it remains difficult for the model to be implemented when large-scale problems are under consideration. Moreover, it does not guarantee that conditional variance H_t is positive definite, and strong restrictions have once more to be imposed on the parameters as well as on the initial variance matrix. Despite its inconsistencies, the VECH model serves as a fundamental benchmark from which the two innate problems of multivariate volatility modelling have been addressed: parsimony and positive definiteness. Finally, the conditional log-likelihood function of the VECH model is

$$L_t(\theta) = -\frac{N(T-1)}{2}\log 2\pi - \frac{1}{2}\log|H_t| - \frac{1}{2}\varepsilon_t' H_t^{-1}\varepsilon_t. \tag{2.40}$$

2.2.2.2. The BEKK model

This model serves as an improved perspective to parameterize H_t and was introduced by Engle and Kroner (1995). In particular,

$$H_t = C_0'C_0 + \sum_{\kappa=1}^{\mathrm{K}} C_{1\kappa}' x_t x_t' C_{1\kappa}' + \sum_{i=1}^{p} A_i \varepsilon_{t-i}\varepsilon_{t-i}' A_i' + \sum_{j=1}^{q} B_j H_{t-j} B_j'. \tag{2.41}$$

The C parameter matrix is restricted to be lower triangular of $\kappa \times \kappa$ dimension, while matrices A, B are left unrestricted of $\kappa \times \kappa$ dimension. The number of the total parameters to be estimated is of order κ^2 or $O(\kappa^2)$. Thus not only there is an apparent reduction of the estimation burden but the model also ensures positive definiteness with no further restrictions to be imposed. However, we still have to restrain κ to take values less than 5. More intuitively, we can rewrite the BEKK model into a VECH equivalent format taking into account that $(ABC) = (C' \otimes A)vec(B)$, where $\otimes$ is a

Kronecker product.

$$H_t = (C_0 \otimes C_0)' vec(I_n) + \sum_{\kappa=1}^{K} (A_{1\kappa} \otimes A_1\kappa)' vec(\varepsilon_{t-1}\varepsilon'_{t-1}) + \\ + \sum_{\kappa=1}^{K} (B_{1\kappa} \otimes B_{1\kappa})' vec(H_{t-1}). \tag{2.42}$$

This implies that VECH and BEKK are equal if and only if

$$C_0 = (C_0 \otimes C_0)' vec(I_n), \tag{2.43}$$

$$A_i = \sum_{\kappa=1}^{K} (A_{i\kappa} \otimes A_i\kappa)', \tag{2.44}$$

$$B_i = \sum_{\kappa=1}^{K} (B_{i\kappa} \otimes B_{i\kappa})'. \tag{2.45}$$

Another modification is when we restrict $A,\ B$ to be diagonal matrices (diagonal BEKK (DBEKK) model). This reduces the number of coefficients even more, since the covariance matrix only depends on its own lagged variance and residuals. However, even after imposing restrictions, both VECH and BEKK alternatives involve a huge number of free parameters to be estimated. To account for this irregularity, Factor and Orthogonal models are suggested instead, which apply common dynamics to all elements of the covariance matrix.

2.2.2.3. Factor Models

Engle, Ng and Rothschild (1990) were the first to observe that asset returns are driven by common underling factors that can be used as future indicators and for parameterization purposes of the conditional-volatility. Later on Harvey, Ruiz and Sentana (1992), Bollerslev and Engle (1993) applied this idea, offering factor models that capture the persistence of conditional variances. A factor model, as Lin (1992) mentions, can be seen as a particular BEKK model and can be defined as follows. Let us assume that we have a BEKK $(1, 1, K)$ model. If $A_\kappa,\ B_\kappa$ have rank one and the same left and right eigenvectors, $\lambda_\kappa,\ w_\kappa\ \forall\kappa = 1, \ldots, K$, then this BEKK model can be seen as Factor GARCH (FGARCH) $(1, 1, K)$ model with:

$$A_\kappa = a_\kappa w_\kappa \lambda'_\kappa, \tag{2.46}$$
$$B_\kappa = \beta_\kappa w_\kappa, \lambda'_\kappa, \tag{2.47}$$

where $a_\kappa,\ \beta_\kappa, w_\kappa$ are $N \times 1$ vectors such that (these are restrictions posed onto the parameters for identification reasons)

$$w'_\kappa \lambda_i = \begin{cases} 0 & for \quad \kappa \neq 1, \\ 1 & for \quad \kappa = i, \end{cases} \tag{2.48}$$

$$\sum_{n=1}^{N} w_{\kappa\eta} = 1. \tag{2.49}$$

The conditional covariance matrix then takes the form

$$H_t = \Omega + \sum_{\kappa=1}^{K} \lambda_\kappa, \lambda'_\kappa (a_\kappa^2 w'_\kappa \varepsilon_{t-1} \varepsilon'_{t-1} w_\kappa + \beta_\kappa^2 w'_\kappa H_{t-1} w_\kappa). \tag{2.50}$$

The vector λ_κ is called the κth factor loading and the scalar $w'_\kappa \varepsilon_t$ is called the κth factor. Several variations have been proposed, e.g. that of Vrontos, Dellaportas, Politis who have presented the Full-Factor GARCH model (2003). The model is characterised by the following equations:

$$y_t = \mu + \varepsilon_t, \tag{2.51}$$
$$\varepsilon_t = W X_t \quad X_t \,|\Phi_{t-1} \sim N_N(0, \Sigma_t), \tag{2.52}$$

where μ is a $N \times 1$ vector of constants, ε_t is a $N \times 1$ innovations vector, W is a $N \times N$ parameter matrix, Φ_{t-1} is the information set at time t-1, $X_t = \{x_{i,t}\}$ is a $N \times 1$ factor vector .Then the conditional covariance matrix

$$H_t = W \Sigma_t W' = W \Sigma_t^{1/2} \Sigma_t^{1/2} W' = \left(W \Sigma_t^{1/2}\right) \left(W \Sigma_t^{1/2}\right)' = LL', \tag{2.53}$$

where $\Sigma_t = diag(h_{1,t}, \ldots, h_{N,t})$ with $h_{i,t} = a_i + b_i x_{i,t-1}^2 + g_i h_{i,t-1}\ \forall i = 1, \ldots, N,\ t = 2, \ldots, T,$ where $h_{i,t}$ is the variance of the ith factor $x_{i,t}$ with $a_i > 0,\ b_i \geq 0,\ g_i \geq 0.$ The factors $x_{i,t}$ are GARCH models while ε_t

represents a vector of linear combinations of these factors. The covariance matrix is always positive definite and the model also accounts for the problem of parsimony. In order to reduce the number of free parameters to be estimated the W has to be restricted to be an identity matrix. Additionally the factors $x_{i,t}$ $\forall i = 1, \ldots, N$, have variances with zero idiosyncratic risk, so there is no need to estimate parameters for them. They are derived through $X_t = W^{-1}\varepsilon_t$. Assuming a multivariate normal distribution for X_t the log-likelihood function is

$$\begin{aligned} L_t(\theta) &= -\frac{(T-1)N}{2}\log 2\pi - \frac{1}{2}\sum_{t=2}^{T}\log|H_t| - \frac{1}{2}\sum_{t=2}^{T}\varepsilon_t' H_t^{-1}\varepsilon_t \\ &= -\frac{(T-1)N}{2}\log 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left[\sum_{i=1}^{N}\log(h_{i,t})\right] - \frac{1}{2}\sum_{t=2}^{T}\left[\sum_{i=1}^{N}\frac{x_{i,t}^2}{h_{i,t}}\right]. \end{aligned} \tag{2.54}$$

The authors propose classical and Bayesian approaches for estimating the parameters. Finally to tackle the problem of model choice Bayesian model averaging using Markov Chain Monte Carlo composition is employed.

2.2.2.4. The Constant Conditional Correlation Model

This is a simple multivariate conditional heteroskedastic model as such it is characterized by Bollerslev (1990). The proposal is to have a model with time-varying conditional variances and covariance but constant conditional correlations. By doing so there is an apparent estimation and specification advantage. The model has been first introduced as an extension to the Seemingly Unrelated Regressions (SUR) model and can be described as follows:

$$y_t = E\left(y_t \,|\psi_{t-1}\right) + \varepsilon_t, \tag{2.55}$$

where y_t is the conditional mean and ψ_{t-1} is the σ-field created through all the available information at time $t-1$.

Let $Var(\varepsilon_t \,|\psi_{t-1}) = H_t$ be the time-varying conditional covariance matrix with h_{ij} elements in H_t and ε_{it} elements in y_t and with conditional correlation

$$h_{ijt} = \rho_{ij}(h_{iit}h_{jjt})^{1/2} \quad \forall i = j+1, \ldots, N, \quad -1 \leq \rho_{ij} \leq 1 \tag{2.56}$$

Then the full conditional covariance matrix is defined as below.

$$H_t = D_t \Gamma D_t, \tag{2.57}$$

with D_t being a stochastic diagonal matrix with standard deviations on the diagonal and zeros elsewhere and Γ a time-invariant matrix of unconditional correlations of the standard residuals both of $\kappa \times \kappa$ dimension. Elements of D_t are simply calculated by univariate GARCH processes.

$$D_t \equiv diag\left(\left[\sqrt{h_{11,t}}, \sqrt{h_{22,t}}, \ldots, \sqrt{h_{\kappa\kappa,t}}\right]\right). \tag{2.58}$$

The log-likelihood function in turn is

$$L_t(\theta) = -\frac{(T-1)N}{2}\log 2\pi - \frac{1}{2}\sum_{t=2}^{T}(\log|D_t\Gamma D_t| - \frac{1}{2}\sum_{t=2}^{T}\varepsilon_t'(D_t\Gamma D_t)^{-1}\varepsilon_t) \tag{2.59}$$

The parameters to be estimated are of order κ^2 or $O(\kappa^2)$ but can be estimated in stages, first the univariate GARCH (p,q) models and then the unconditional correlation matrix, which both are very simple to compute. The difference with the BEKK model is that it requires simultaneous parameters estimation, which can become extremely arduous for higher-dimensional matrices. Moreover, the model ensures positive definiteness and parsimony, though the constant correlation assumption is too strong.

2.2.2.5. The Dynamic Conditional Correlation Model

This model's attainment is to amalgamate the flexibility of estimation of the univariate GARCH models for the covariance matrices with parsimonious parametric models for the conditional correlation matrix. Engle (2002) is motivated by the need to provide more reliable multivariate volatility estimates and uses the following decomposition of the covariance matrix:

$$r_t|F_{t-1} \quad \sim \quad N(0, D_t R_t D_t),$$

$$
\begin{aligned}
H_t &= D_t R_t D_t, && (2.60)\\
\varepsilon_t &= D_t^{-1} r_t, && (2.61)\\
R_t &= Q_t^{*-1} Q_t Q_t^{*-1}, && (2.62)
\end{aligned}
$$

where $Q_t = S \circ (\iota\iota' - A - B) + A \circ \varepsilon_{t-1}\varepsilon'_{t-1} + B \circ Q_{t-1}$, with ι a vector of units and $\circ$ to be the Hadamard product which multiplies element by element two equal-sized matrices, and $Q_t^* = diag\left[\sqrt{q_{11,t}, q_{22,t}, \ldots, q_{\kappa\kappa,t}}\right]$. Q_t can be rewritten as $Q_t = (1-\alpha-\beta)\bar{Q}_t + a\varepsilon_{t-1}\varepsilon'_{t-1} + \beta Q_{t-1}$ where $\bar{Q}$ is the unconditional correlation of standardized residual ε_t. The log-likelihood to estimate the parameters θ is

$$
\begin{aligned}
L_t(\theta) &= -\frac{(T-1)N}{2}\log 2\pi - \frac{1}{2}\sum_{t=2}^{T}(\log|D_t R_t D_t|) - \frac{1}{2}\sum_{t=2}^{T} r'_t (D_t^{-1} R_t^{-1} D_t^{-1}) r_t \\
&= -\frac{(T-1)N}{2}\log 2\pi - \frac{1}{2}\sum_{t=2}^{T}(\log|D_t| + r_t D_t^{-1} D_t^{-1} r_t - \varepsilon'_t\varepsilon_t + \log|R_t| + \\
&+ \varepsilon'_t R_t^{-1}\varepsilon_t).
\end{aligned}
\tag{2.63}
$$

This model has the advantage (similar to CCC) that its parameters can be estimated in stages. First the univariate GARCH(p,q) models are fitted and estimates of h_{it} are obtained; then the conditional correlation matrix can be calculated. In total the parameters to be calibrated are 3κ for the covariance matrix and $2 + \frac{\kappa(\kappa+1)}{2}$ for the correlation matrix. The model is always parsimonious and ensures a positive-definite covariance matrix estimate. This also has as a consequence that it can be used for large κ-dimensional progressions.

An alternative to the DCC model of Engle is proposed by Tse and Tsui (2002). Their model, the Multivariate GARCH with time-varying correlations (VC-MGARCH), can be defined as below:

$$
\begin{aligned}
H_t &= D_t R_t D_t, && (2.64)\\
R_t &= (1-\theta_1-\theta_2)R + \theta_1\Psi_{t-1} + \theta_2 R_{t-1}, && (2.65)
\end{aligned}
$$

where $\theta_1,\ \theta_2$ are nonnegative and $\theta_1+\theta_2 \leq 1$, $R = \{\varrho_{ij}\}$ is the unconditional

symmetric $\kappa \times \kappa$ positive definite correlation matrix and Ψ_{t-1} is the $\kappa \times \kappa$ matrix whose elements are functions of the lagged returns. If Ψ_{t-1} and Γ_0 are well-defined correlation matrices, then Γ_t will also be a correlation matrix. In addition, $\Psi_{t-1}\{\psi_{ij,t-1}\}$ is the matrix of the sample correlation of the standardized residuals $\{\epsilon_{t-1}, \ldots, \epsilon_{t-M}\}$, where

$$\psi_{ij,t-1} = \frac{\sum_{m=1}^{M} \epsilon_{i,t-m}\epsilon_{j,t-m}}{\sqrt{\sum_{m=1}^{M} \epsilon_{i,t-m}^2 \sum_{m=1}^{M} \epsilon_{j,t-m}^2}}, \; 1 \leq i < j \leq \kappa, \tag{2.66}$$

$\epsilon_{i,t} = \frac{\varepsilon_{i,t}}{\sqrt{h_{iit}}}$ and the condition $M \geq \kappa$ is necessary in order to ensure that Ψ_{t-1} is positive definite. Finally the log-likelihood is formulated as below:

$$\begin{aligned} L_t(\theta) &= -\frac{1}{2}\log|D_t R_t D_t| - \frac{1}{2} r_t' D_t^{-1} R_t^{-1} D_t^{-1} r_t \\ &= -\frac{1}{2}\log|R_t| - \frac{1}{2}\sum_{i=1}^{\kappa} \log h_{it} - \frac{1}{2} r_t' D_t^{-1} R_t^{-1} D_t^{-1} r_t \end{aligned} \tag{2.67}$$

2.2.2.6. The Exponential Smother

This is the multivariate alternative of the Exponential Moving Average model or Risk Metrics. The approach employed to model the covariance matrix is

$$H_t = \lambda H_{t-1} + (1-\lambda)\varepsilon_{t-1}\varepsilon_{t-1}', \tag{2.68}$$

where again as in the univariate model λ equals 0.94 for daily and 0.97 for monthly data; thus there is no need to estimate parameters. The model guarantees parsimony and positive definiteness for the covariance matrix if H_0 is chosen to be a positive-definite matrix as well. Each individual element equals $h_{ijt} = \lambda h_{ijt-1} + (1-\lambda)\varepsilon_{it-1}\varepsilon_{jt-1}$, which in turn implies that this model is a restricted VECH model where C equals 0, A, B sum to one and λ is the weight of each individual variance and covariance.

2.2.2.7. Orthogonal GARCH

What has originally been suggested by Kariya (1988) and afterwards by Alexander and Chibumba (1997) is to take as system X_t of κ assets or risk

factors that are unconditionally, uncorrelated where,

$$X_t = \hat{\Sigma}^{-1/2}(r_t - \mu_t), \tag{2.69}$$

In the matrix $X_t\{x_i\}\ \forall i = 1, \ldots, k$ the components have been standardized such that $x_i = \frac{y_i - \mu_i}{\sigma_i}$. If we additionally impose that the variables in X_t are also conditionally uncorrelated, then their conditional covariance matrix will be diagonal.

$$\begin{aligned}
X_t &= \hat{\Sigma}^{-1/2}\varepsilon_t, && (2.70)\\
X_{t+1}|F_t &\sim N(0, H^x_{t+1}), && (2.71)\\
H^x_{t+1} &= diag[h^x_{11,t+1}, h^x_{22,t+1}, \ldots, h^x_{kk,t+1}], && (2.72)\\
h^x_{ii,t+1} &= \omega_i + \beta_i h^x_{ii,t} + \alpha_i([\hat{\Sigma}^{-1/2}\varepsilon_t]_{[i]})^2, && (2.73)\\
H^r_{t+1} &= \hat{\Sigma'}^{-1/2} H^x_{t+1} \hat{\Sigma}^{-1/2}. && (2.74)
\end{aligned}$$

At the end we are left with $\frac{\kappa(\kappa-1)}{2} + 3\kappa$ parameters to be estimated. The model offers parsimony and ensures positive definite conditional covariance matrix, but it does not fit the data when there is no evidence of strong correlation or when negatively correlated variables exist.

2.2.2.8. Generalized Orthogonal GARCH

This is a natural generalization of O-GARCH and it is proposed by Van Der Weide (2002). Here the assumption of orthogonality of components is relaxed and linkages between any potential invertible matrixes are created. The general assumption of the model is that the observed economic process is governed by a linear combination of uncorrelated economic components

$$x_t = Z y_t, \tag{2.75}$$

where Z is a linear map, invertible and constant over time while it represents the relationships between observed variables and unobserved components. These unobserved components are assumed to have zero mean and unit variance, hence the unconditional covariance matrix equals

$$V = E x_t x_t^T = Z Z^T. \tag{2.76}$$

The conditional covariance matrix is given by

$$V = ZH_tZ^T. \tag{2.77}$$

If we assume this Z linear map, then there exists an orthogonal matrix U_0 such that,

$$P\Lambda^{1/2}U_0 = Z, \tag{2.78}$$

where P is orthogonal eigenvectors matrix and Λ the orthogonal eigenvalues matrix. Here U_0 is an orthogonal matrix. To communicate more clearly the way U_0 is expressed, we should provide further clarification.

If $U_0 = U$, then it can be proved that every m-dimensional orthogonal matrix U with $det(U) = 1$ can be represented as a product of $\begin{pmatrix} m \\ 2 \end{pmatrix} = \frac{[m(m-1)]}{2}$ rotation matrices such that $U = \prod_{i<j} R_{ij}(\theta_{ij})$ with $-\pi \leq \theta_{ij} \leq \pi$ and θ_{ij} the angle by which the rotation $R_{ij}(\theta_{ij})$ is performed from e_i to e_j (it can be estimated using maximum likelihood). As a consequence, U is expressed as the product of rotation matrices, while in O-GARCH $U = I_N$.

The implied correlation here can be computed as time-varying processes, and equals

$$R_t = D_t^{-1}V_tD_t^{-1}, \tag{2.79}$$

where

$$D_t = (V_t \circ I_m)^{1/2} \tag{2.80}$$

and

$$V_t = ZH_tZ^T \tag{2.81}$$

is the conditional covariance matrix of $x_t = Zy_t$. The parameters to be estimated include the rotation coefficients and the parameters for the univariate GARCH models. The log-likelihood of the model can be expressed as

$$\begin{aligned} L_t(\theta, \alpha, \beta) &= -\frac{(T-1)N}{2}\log 2\pi - \frac{1}{2}\sum_{t=2}^{T}(\log\left|ZH_tZ^T\right|) - \dots \\ &- \frac{1}{2}\sum_{t=2}^{T} y_t^T Z^T (ZH_tZ^T)^{-1} Zy_t \end{aligned}$$

$$= -\frac{(T-1)N}{2}\log 2\pi - \frac{1}{2}\sum_{t=2}^{T}\Big(\log\Big|ZZ^T\Big| + log\,|H_t| + y_t^T H_t^{-1} y_t\Big). \tag{2.82}$$

It is substantiated that since orthogonal GARCH models are particular F-GARCH models, more general BEKK models had GO-GARCH models nested. Thus for the MLE to be consistent and the process to be covariance stationary the parameters α_i, β_i should also be stationary, thus $\alpha_i + \beta_i < 0\ \forall i = 1, \ldots, m$

2.2.2.9. Asymmetric Multivariate GARCH Models

A natural extension to the above-discussed multivariate models is to incorporated attributes which have been successfully captured by univariate volatility schemes (leverage effects etc.).

2.2.2.10. General Dynamic Covariance Matrix model

Kroner and Ng (1998) had an intuitive motivation to try to overcome strong restrictions imposed by some widely used multivariate volatility models and through robust conditional moments to try to account for any misspecifications that might be detected. For this reason they nested VECH, BEKK, CCC and Factor GARCH model while they add an extension which can incorporate asymmetric dynamics affecting volatility. The definition first of the General Dynamic Covariance Matrix model (GDC) follows.

$$\begin{aligned}
H_t &= D_t R D_t + \Phi \odot \Theta_t, && (2.83)\\
D_t &= [d_{ijt}], && (2.84)\\
d_{iit} &= \sqrt{\theta_{iit}}\ \forall i \quad d_{ijt} = 0\ \forall i \neq j, && (2.85)\\
\Theta_t &= [\theta_{ijt}], && (2.86)\\
\theta_{ijt} &= \omega_{ij} + b_i' H_{t-1} b_j + a_i' \varepsilon_{t-1} \varepsilon_{t-1}' a_j\ \forall i, j, && (2.87)\\
R &= [r_{ij}] \quad r_{ii} = 0\ \forall i, && (2.88)\\
\Phi &= [\phi_{ij}] \quad \phi_{ii} = 0\ \forall i, && (2.89)
\end{aligned}$$

where $a_i, b_i\ \forall i = 1, \ldots, \kappa$ are $\kappa \times 1$ vectors κ being the number of variables) and $\omega_{ij},\ \rho_{ij},\ \phi_{ij}\ \forall i, j = 1, \ldots, \kappa$ are scalars with $\Omega \equiv [\omega_{ij}]$ positive definite.

The GDC model is comprised of two parts. The first is a constant correlation model with variances given by a BEKK model, and the second part is $\Phi \odot \Theta_t$, a matrix with zero diagonal elements and non-diagonal elements given by BEKK type covariance functions scaled by the ϕ_{ij} parameters ($h_{iit} = \theta_{iit}\ \forall i\ h_{ijt} = \rho_{ij}\sqrt{\theta_{iit}}\sqrt{\theta_{jjt}} + \phi_{ij}\theta_{ijt} \quad \forall i \neq j$). Different restrictions imposed on parameters can yield a restricted VECH, CCC, BEKK or FARCH. This can be used in order to compare different existing models and has been the ground on which extensions were introduced, such as the Asymmetric Dynamic Covariance Matrix model (ADC).

2.2.2.11. Asymmetric Dynamic Covariance Matrix model

This extension (Kroner and Ng (1998)) nests the previous generalized model with the asymmetric impact of news to the returns. Let us initially define a multivariate vector of returns:

$$\begin{aligned} r_t &= \mu_t + \varepsilon_t, \\ \varepsilon_t | F_t &\sim N(0, H_t). \end{aligned} \tag{2.90}$$

If we define η_t as

$$\begin{aligned} \eta_{it} &= max[0, -\varepsilon_{it}], \\ n_t &= [\eta_{1t}, \ldots, \eta_{\kappa t}], \end{aligned} \tag{2.91}$$

then an ADC model can be expressed as follows:

$$H_t = D_t R D_t + \Phi \odot \Theta_t, \tag{2.92}$$
$$D_t = [d_{ijt}], \tag{2.93}$$
$$d_{iit} = \sqrt{\theta_{iit}}\ \forall i \quad d_{ijt} = 0\ \forall i \neq j, \tag{2.94}$$
$$\Theta_t = [\theta_{ijt}], \tag{2.95}$$
$$\theta_{ijt} = \omega_{ij} + b_i' H_{t-1} b_j + a_i' \varepsilon_{t-1} \varepsilon_{t-1}' a_j + g_i' \eta_{t-1} \eta_{t-1}' g_j\ \forall i, j, \tag{2.96}$$
$$R = [r_{ij}] \quad r_{ii} = 1\ \forall i \quad r_{ij} = \rho_{ij}\ \forall i \neq j, \tag{2.97}$$
$$\Phi = [\phi_{ij}], \quad \phi_{ii} = 0 \quad \forall i. \tag{2.98}$$

The same hold for $a_i,\ b_i,\ \omega_{ij},\ \rho_{ij},\ \phi_{ij}$ and Ω as for GDC model. The

number of parameters to be estimated is $\frac{9\kappa^2}{2} + \frac{3\kappa}{2}$. The difference between those two models is in the $g_i'\eta_{t-1}\eta_{t-1}'g_j$ component that is added in the θ_{ijt} equation. Thus not only does the ADC nest some natural extension of those four models discussed before, but it also incorporates asymmetric effects governing volatility estimation. Under a set of conditions the model takes the form of one of the four extended models presented shortly below, and gives estimates regarding the covariance matrix accordingly.

2.2.2.12. Asymmetric VECH

$$h_{iit} = \omega_{ii} + \beta_i^2 h_{iit-1} + a_i^2\varepsilon_{it-1}^2 + \gamma_i^2\eta_{it-1}^2 \quad \forall i, \tag{2.99}$$

$$h_{ijt} = \phi_{ij}\omega_{ii} + \phi_{ij}\beta_i\beta_j h_{ijt} + \phi_{ij}a_i a_j\varepsilon_{it-1}\varepsilon_{jt-1} + \phi_{ij}\gamma_i\gamma_j\eta_{it-1}\eta_{jt-1} \quad \forall i \neq j. \tag{2.100}$$

2.2.2.13. Asymmetric CCC

$$h_{iit} = \omega_{ii} + \beta_i^2 h_{iit-1} + a_i^2\varepsilon_{it-1}^2 + \gamma_i^2\eta_{it-1}^2 \quad \forall i, \tag{2.101}$$

$$h_{ijt} = \rho_{ij}\sqrt{h_{iit}}\sqrt{h_{jjt}} \quad \forall i \neq j. \tag{2.102}$$

2.2.2.14. Asymmetric BEKK

$$H_t = \Omega + A'\varepsilon_{t-1}\varepsilon_{t-1}'A + B'H_{t-1}B + G'\eta_{t-1}\eta_{t-1}'G. \tag{2.103}$$

2.2.2.15. Asymmetric FARCH

$$h_{iit} = \sigma_{ij} + \lambda_i\lambda_j h_{pt} \quad \forall i, \tag{2.104}$$

$$h_{pt} = \omega_p + \beta h_{pt-1} + a\varepsilon_{pt-1}^2 + \gamma\eta_{pt-1}^2 \quad \forall i \neq j, \tag{2.105}$$

$$h_{pt} \equiv w'H_t w, \tag{2.106}$$

$$\varepsilon_{pt} \equiv w'\varepsilon_t, \tag{2.107}$$

$$\eta_{pt} \equiv w'\eta_t, \tag{2.108}$$

$$\sigma_{ij} \equiv \omega_{ij} - \lambda_i\lambda_j w'\Omega w. \tag{2.109}$$

Applications of the models can be found in Bekaert and Wu (2000), Brooks and Henry (2000) and finally Isakov and Perignon (2000).

2.2.2.16. Asymmetric Dynamic Conditional Correlation

Cappiello, Engle and Shephard (2006) offer an extended version of the DCC model (Engle, 2002) so as to account for asymmetries in the conditional variances, covariances and correlations. Specifically, considerable attention is drawn to the impact of negative returns to correlations and covariances, but also an endeavor is made to investigate whether these effects are apparent not only for stocks but for fixed income securities also . As in DCC returns are summed to be normal with zero mean while the remaining driving dynamics of the model are given from the following equations.

$$
\begin{aligned}
r_t &= \mu_t + \varepsilon_t, && (2.110)\\
\varepsilon_t|F_{t-1} &\sim N(0, H_t),\\
H_t &= D_t R_t D_t, && (2.111)\\
R_t &= Q_t^{*-1} Q_t Q_t^{*-1}, && (2.112)\\
Q_t &= \bar{Q} - A'\bar{Q}A - B'\bar{Q}B - G'\bar{Q}G + A'\nu_{t-1}\nu_{t-1}'A + B'Q_{t-1}B + ...\\
&\quad + G'\eta_{t-1}\eta_{t-1}'G, && (2.113)\\
\nu_t &= D_t^{-1}\varepsilon_t, && (2.114)\\
\eta_t &= \varepsilon_t \cdot \mathbf{1}\{\varepsilon_t < 0\}, && (2.115)\\
\bar{N} &= E\left[\eta_t, \eta_t'\right], && (2.116)\\
Q_t^* &= diag\left[\sqrt{q_{11,t}, q_{22,t}, \ldots, q_{\kappa\kappa,t}}\right], && (2.117)
\end{aligned}
$$

where $q_{ii,t}$ is the $(i,i)^{th}$ element of Q_t. Where expectations of $\bar{N}$, $\bar{Q}$ are difficult to find they are substituted by $\frac{1}{T}\sum_{t=1}^{T}\eta_t\eta_t'$ and $\frac{1}{T}\sum_{t=1}^{T}\nu_t\nu_t'$ respectively. Normally the element $\rho_{ij,t}$ of R_t equals $\rho_{ij,t} = \frac{q_{ij,t}}{\sqrt{q_{ii,t}q_{jj,t}}}$.

2.2.2.17. Matrix Exponential GARCH

The introduction of this model is very recent, from Kawakatsu (2006). This is a multivariate approach of EGARCH taking advantage of properties of the matrix exponential which ensures positive definiteness without further constrains. The model also incorporates leverage while allowing a variety of distributions for estimation purposes. The specification of the model follows. Let $\log H_t$ depend on its own past and lagged innovations; then

$$\bar{h}_t = \sum_{i=1}^{p} A_i \bar{h}_{t-i} + \sum_{j=1}^{q} B_j \varepsilon_{t-j} + \sum_{j=1}^{q} F_j \left(|\varepsilon_{t-j}| - E\left[|\varepsilon_{t-j}| \right] \right), \tag{2.118}$$

where $\bar{h}_t$ is a $\frac{\kappa(\kappa+1)}{2}$ matrix and $\bar{h}_t = h_t - c_0,\ h_t = vech(\log H_t),\ c_0 = vech(C)$, C is a symmetric $\kappa \times \kappa$ matrix and $c_0,\ A_i,\ B_j,\ F_j$ are parameter matrices of dimension $\kappa^* \times 1$, $\kappa^* \times \kappa^*$, $\kappa^* \times \kappa$, $\kappa^* \times \kappa$ respectively. The B_j parameter captures the leverage effect while there is no need to impose any restrictions since $\log H_t$ is positive definite. The number of parameters to be estimated is of order κ^4, $O\left(\kappa^4\right)$, which is a considerable number; thus the author restricts κ to equal 2 or 3. This implies that although positive definiteness is achieved, parsimony is still questionable; thus diagonal restrictions have to be imposed on parameters to reduce the number of free parameters.

2.2.2.18. Modelling Multivariate Volatility with Copula-GARCH Models

Let us define two variables U=F(X) and V=G(Y) and a copula H which is their joint distribution and completely captures their dependence. Note that this copula has all dependence information but individual information about F, G densities. Sklar (1959) proposes that any N-dimensional joint distributions may be decomposed into N marginal distributions and a copula. Patton (2000) verifies that SKlar's theorem can be extended to conditional distributions so as to model time-varying conditional distribution. The idea in this model-category is to estimate conditional variances using univariate GARCH models, set the marginal distribution and compute the correlation using copula functions. An example of this approach can be found also in Jondeau, Rockinger (2001).

2.2.2.19. Long Memory Volatility Models

Denote a time series by $\{y_t\}_{t=0}^{T}$ and its correlation by $\rho(\kappa)$. Engle and Bollerslev (1986), Taylor (1986) observed the persistence of volatility (long memory). In general a time series is said to have long memory if "it has a slowly declining autocorrelation function (ACF) and an infinite spectrum at zero frequency" (Granger and Ding (1996)). In other words a stationary

series with long memory displays autocorrelation coefficients that decline slowly at a hyperbolic rate while in particular long memory in volatility implies that the effects of shocks perish slowly. A simple illustration of long memory series is a parametric fractional integrated noise process $I(d)$.

$$
\begin{aligned}
(1-L)^d y_t &= \eta_t \quad \eta_t \sim iid(0,\sigma^2), && (2.119)\\
(1-L)^d &= \sum_{\kappa=0}^{\infty} \psi_\kappa L^\kappa, && (2.120)\\
\psi_0 &= 1,\ \psi_\kappa = \prod_{j=1}^{\kappa} (1-\frac{1+d}{j}), && (2.121)
\end{aligned}
$$

where L is the lag operator and d is called the fractional degree of integration of a process and $(1-L)^d$ is the fractional operator. This is the workhorse behind ARFIMA and FIGARCH, models specifically constructed to account for autocorrelation persistence in returns when trying to forecast volatility. In order to measure the degree of long-memory existence in a time series, instead of the theoretical ACF with sample quantities, there have been proposed semi parametric and non-parametric tests and estimators.

2.2.2.20. Fractionally Integrated GARCH (FIGARCH)

Since Taylor (1986) first observed that the absolute values of returns have slow decaying autocorrelation functions the use of historical volatility models as well as ARCH models when attempting fractional integration has become an article of faith among researchers. Financial applications include these of Bailie (1996), Bollerslev and Mikkelsen (1996), who have introduced FIGARCH while integrating squared residuals of US dollar-Deutschemark exchange rates of conditional-volatility. Later Bollerslev and Mikkel use the FIEGARCH (1996, 1999) as an alternative to study S&P500 volatility, as also Taylor(2000) does. Vilasuso tests FIGARCH against GARCH and IGARCH (2002), while Li (2002), Martens and Zein (2004), Pong, Shackleton, Taylor and Xu (2004) compare FIGARCH with implied volatility. Recall also that Hwang and Satchell (1998) study long-ARFIMA as Li (2002) does, while Bollerslev, Diebold and Andersen (2003) propose a Vector Autorgrassive Realised Volatility model with long lags (VAR-RV).

If we parametrised the conditional variance of a GARCH (p,q) model,

then we have

$$h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}^2 \equiv \omega + a(L)\varepsilon_t^2 + \beta(L)h_t, \tag{2.122}$$

where L is the lag or "backshift operator," which in turn implies a prerequisite that the lag polynomial $[1-\beta(L)]^{-1}a(L)$ must be positive. Rearranging the above equation, we get [1-a (L) -$\beta(L)]\varepsilon_t^2 = \omega + [1 - \beta(L)]\nu_t]$, where $\nu_t \equiv \varepsilon_t^2 - h_t \Rightarrow E_{t-1}(\nu_t) = 0$. Engle and Bollerslev (1986) proposed the Integrated GARCH (IGARCH), where $1 - \hat{\beta}(x) - \hat{a}(x) \equiv (1 - L)\phi(L) = 0$. The model can be rewritten as $\phi(L)(1 - L)\varepsilon_t^2 = \omega + [1 - \beta(L)]\nu_t$. Empirical evidence rejects the IGARCH model, although it is observable that in general the volatility is a mean-reverting process. In order to address this inconsistency Bailie, Bollerslev and Mikkelsen (1996) established the FIGARCH model of order (p,d,q). FIGARCH is formulated as follows:

$$\phi(L)(1 - L)^d \varepsilon_t^2 = \omega + [1 - \beta(L)]\nu_t. \tag{2.123}$$

Setting $\omega = \phi(L)(1-L)^d(\varepsilon_t^2 - \sigma^2)$ we can have a more explicit expression of FIGARCH model as the one:

$$[1 - \beta(L)]h_t = -\beta(L)\varepsilon_t^2 + \varepsilon_t^2 - \phi(L)(1 - L)^d(\varepsilon_t^2 - \sigma^2). \tag{2.124}$$

The FIGARCH model reduces to the covariance stationary GARCH (p,q) for d=0 and the IGARCH(p,q) for d=1, whereas the roots of the polynomial $\phi(z) = 0$ lie outside the unit circle. Although the model is not weakly stationary, by restricting d to vary between 0 and 1 ($0 \leq d < 1$) we have a strictly stationary process, and we add flexibility to the estimation of long-term dependence of the conditional variance.

Additionally let us assume a Fractionally Integrated ARMA process for the mean, in addition to the FIGARCH assumption for the variance, where the short-run characteristics of the time series are captured by the conventional ARMA parameters and the long-run attributes are modeled through the fractional differential parameter.

$$\psi(L)(1 - L)^{d_o}(y_t - \mu) = \theta(L)\varepsilon_t, \tag{2.125}$$

where $\psi(L) = 1 - \sum_{j=1}^{\alpha} \psi_j L^j$, d_o with $-0.5 \leq d_o < 0.5$ a fraction number that allows y_t to exhibit long memory and $\theta(L) = 1 - \sum_{j=1}^{m} \theta_j L^j$ with α, m defining the order of the polynomials for the two equations respectively and $(1-L)^{d_o} = \sum_{j=0}^{\infty} \frac{\Gamma(j-d_o)}{\Gamma(j+1)\Gamma(-d_o)} L^j \equiv \sum_{j=0}^{\infty} \pi_j(d_o) L^j$ with $\pi_j(z) \equiv \Gamma(j-z)/\Gamma(j+1)\Gamma(-z)$. In accordance with the above mentioned, FIGARCH(1,d,1) model can be rewritten as

$$\begin{aligned} h_t &= \omega + [1 - (1-\beta L)^{-1}(1-\phi L)(1-L)^d]\varepsilon_t^2 \\ &\equiv \omega + \lambda(L)\varepsilon_t^2, \qquad (2.126) \\ \omega &= (1-\beta L)^{-1}(1-\phi L)(1-L)^d \sigma^2, \qquad (2.127) \end{aligned}$$

where according to Bollerslev and Mikkelsen (1996),

$$\begin{aligned} \lambda(L) &= \sum_{j=1}^{\infty} \lambda_j L^j, \qquad (2.128) \\ \lambda_1 &= d - \beta + \phi, \qquad (2.129) \\ \lambda_2 &= (d-\beta)(\beta-\phi) + \frac{d(d-1)}{2}, \qquad (2.130) \\ \lambda_\kappa &= \beta^{k-2}(d-\beta)(\beta-\phi) - (\beta-\phi)\sum_{j=2}^{k-1} \pi^j(d)\beta^{k-1-j} - \pi_k(d) \quad \forall k = 3, 4, \ldots \\ & \qquad (2.131) \end{aligned}$$

ω, λ are non negative to ensure positive variance. The model is not weakly stationary but when $0 \leq d_o < 1$ can be strictly stationary.

2.2.2.21. Fractionally Intergraded EGARCH (FIEGARCH)

When EGARCH has been proposed in order to account for leverage effects by Nelson (1991), the possibility of the roots of the polynomial being close to one (thus the model to exhibit unit root) had not been addressed adequately. Thus an extension is imperative to allow for fractional orders of integration. First the classical EGARCH model is rewritten as below and the FIEGARCH is introduced.

$$\begin{aligned}\ln(h_t) &= \omega + (1 - \sum_{i=1,p} \phi_i L^i)^{-1}(1 + \sum_{i=1,p} \psi_i L^i) g(z_{t-1}) \\ &\equiv \omega + [1 - \phi(L)]^{-1}[1 + \psi(L)] g(z_{t-1}), \end{aligned} \tag{2.132}$$

where $g(z_t) = \theta z_t + \gamma[|z_t|] - E(|z_t|)$. θ measures the effect of different signed shocks on volatility. Thus if $\theta < 0$ the future conditional variance will increase disproportionately more following bad news than good news. If we now factorize the $[1-\phi(L)] = \phi(L)(1-L)^d$, we get the FIEGARCH (p,d,q):

$$\ln(\sigma_t^2) = \omega + \phi(L)^{-1}(1 - L)^{-d}[1 + \psi(L)] g(z_{t-1}) \tag{2.133}$$

The FIGARCH model combines the EGARCH for d=0 and the IGARCH (p,q) for d=1, whereas the roots of the polynomial $\phi(z) = 0$, as with FIGARCH, lie outside the unit circle. Restrictions again have to be imposed for covariance stationarity on d, thus d should lie between $-0.5 < d < 0.5$. In contrast to the FIGARCH model there is no extra need to limit parameters to be positive.

2.3. Stochastic Volatility Models

2.3.1. Univariate Stochastic Volatility Models

In the original Black Scholes model there are two assets. The first is a riskless asset (bond), with price B_t at time t following the ordinary differential equation

$$\mathrm{dB_t} = \mathrm{rB_t dt}, \tag{2.134}$$

where r is the spot interest rate at time t and it is a non-negative constant. The second asset is a risky stock S_t that is described according to the stochastic differential equation

$$\mathrm{dS_t} = \alpha \mathrm{S_t dt} + \sigma S_t \mathrm{dW_t}, \tag{2.135}$$

where α is a constant mean return rate or else a drift, $\sigma>0$ is a constant volatility parameter and W_t is a geometric Brownian motion which also represents the random source of the model. The Black-Scholes model relies in

a number of assumptions that are inherently counterfactual. Outlining the most important of them, we find the presupposition of a continuity for the stock price process (no jumps and a log normal distribution at expiry), no transaction costs are imposed and no arbitrage opportunities exist (this in tern implies that the market is frictionless), continuous trading and short selling with full use of proceeds is possible and finally the volatility is constant. By relaxing the last assumption, stochastic volatility models emerged allowing the second moment to change randomly according to some stochastic differential equation or some discrete random process. These models can also be used to explain the discrepancies between Black-Scholes predicted and those observed option prices (smile curves) using implied volatility. Implied volatility is the value of volatility that must go into the Black-Scholes formula to yield the observed option price, and is involved in a variety of financial applications since the actual volatility process is not directly observable and can be approximated using proxies.

A pure stochastic volatility process is as follows:

$$\mathrm{dS_t} = \mu \mathrm{S_t dt} + \sigma_t S_t \mathrm{dW_t}, \tag{2.136}$$

where σ_t changes stochastically according to time and can be for example an *Itô* drift diffusion process, a jump process etc. Moreover it is apparent that the volatility has itself a random source and thus there are no unique martingales to measure the process. Therefore to complete the above pricing formula we additionally denote

$$\sigma_t = f(V_t), \tag{2.137}$$

$$dV_t = (\mu + \beta V_t)dt + \gamma V_t dZ_t. \tag{2.138}$$

2.3.1.1. Mean-Reverting Stochastic Volatility models

Researchers dealing with stochastic volatility have always favoured the idea of mean reversion, which refers to a linear retract term in the drift term or in the underlying process that governs the volatility. When a model mean reverts, we usually say that it follows an Ornstein-Uhlenbeck process. Most of the models discussed here are simple one-factor models initially used to capture interest-rate movements. However they can very easily be extended

to be used to describe the volatility evolution.

2.3.1.2. The Vasicek Model

Vasicek (1977) is one of the first to suggest a mean-reverting stochastic-volatility model with constant coefficients. The dynamics that drive Vasicek's model can be represented as follows:

$$dS_t = \mu S_t dt + \sigma_t S_t dW_t, \tag{2.139}$$
$$dV_t = \eta(\theta - V_t)dt + \gamma dZ_t, \tag{2.140}$$

where η is the rate of mean reversion and θ the long-term mean.

2.3.1.3. The Cox-Ingersoll-Ross (CIR) model

Another popular mean-reverting model for stochastic volatility which is introduced as an extension of the Vasicek model is the Cox, Ingersoll and Ross model, named after those that suggested it (1985). The model can be defined by the following equations:

$$dS_t = \mu S_t dt + \sigma_t S_t dW_t, \tag{2.141}$$
$$dV_t = \eta(\theta - V_t)dt + \gamma\sqrt{V_t}dZ_t, \tag{2.142}$$

where $W_t,\ Z_t$ are uncorrelated Winner processes. The $\gamma\sqrt{V_t}$ term ensures the positivity of $V_t\ \forall\eta, \theta > 0$, while by further imposing $2\eta\theta > \gamma^2$ we exclude zero values for V_t as well.

2.3.1.4. Heston's model

What was unique when Heston first proposed his model in 1993 was that the vast majority of all researchers prior to that date assumed a zero risk premium or zero correlation coefficient for the stochastic-volatility process. This model allows for arbitrary correlation between volatility and returns as well as a stochastic process for interest rates and a closed-form solution for bond and currency option prices. The model first assumes that the spot asset has the following diffusion process:

$$dS_t = \mu S_t dt + \sqrt{V_t} S_t dZ_{1t}. \tag{2.143}$$

If the volatility follows a CIR process, then

$$dV_t = \eta(\theta - V_t)dt + \gamma\sqrt{V_t} dZ_{2t}, \tag{2.144}$$

where $Corr(dZ_{1t},\ dZ_{2t}) = \rho$, with $Z_{1t},\ Z_{2t}$ Wiener processes.

2.3.2. Multivariate Stochastic-Volatility Models

Since the first univariate stochastic-volatility model was proposed by Taylor (1982, 1986), a vast literature has been published covering different aspects of the same notion, that volatility evolves over time following a stochastic diffusion process. Multivariate SV though has been discussed a decade later from Harvey *et al.* in 1994. In general if we consider a vector which includes the log prices of m assets, $S = (S_1, \ldots, S_m)'$ and a y vector of their returns $y = (y_1, \ldots, y_m)'$, then the general diffusion model for S is

$$\begin{aligned} dS_t &= H_t^{1/2} dW_{1t}, & (2.145)\\ df(vech(H_t)) &= \alpha vech(H_t)dt + \beta vech(H_t)dW_{2t}, & (2.146) \end{aligned}$$

where $W_{1t},\ W_{2t}$ are vectors of Brownian motions and $df(vech(H_t))$ is the instantaneous covariance matrix. By following the Euler approximation we end with the following discrete MSV model:

$$\begin{aligned} y_t &= H_t^{1/2}\varepsilon_1 \quad \varepsilon_t \sim N(0, I_m), & (2.147)\\ f(vech(H_t)) &= \alpha vech(H_{t-1}) + f(vech(H_{t-1})) + \beta vech(H_{t-1})\eta_{t-1}, & (2.148)\\ \eta_{t-1} &\sim N(0, \Sigma_n). \end{aligned}$$

The basic intricacy of the above equation is that parameters have to be restricted for the covariance matrix to be positive definite and parsimonious.

2.3.2.1. Introducing the fundamental model

As mentioned before, Harvey *et al.* (1994) specify the first MSV model:

$$y_t = H_t^{1/2}\varepsilon_t, \tag{2.149}$$

$$H_t^{1/2} = diag\left\{\exp\left(\frac{h_{1t}}{2}\right),\ldots,\exp\left(\frac{h_{mt}}{2}\right)\right\} = diag\left\{\exp\left(\frac{h_t}{2}\right)\right\}, \tag{2.150}$$

$$h_{1+t} = \mu + \phi \circ h_t + \eta_t, \tag{2.151}$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathrm{P}_\varepsilon & 0 \\ 0 & \Sigma_n \end{pmatrix}\right], \tag{2.152}$$

with $\Sigma_n = \{\sigma_{n,ij}\}$ a positive definite covariance matrix and $\mathrm{P}_\varepsilon = \{\rho_{ij}\}$ the correlation matrix with $\rho_0 = 1$ and $|\rho_{ij}| < 1\ \forall i \neq j,\ i,j = 1,\ldots,m$ and $\circ$ the element by element multiplication factor . Since H_t is a diagonal matrix we ensure it is positive definite and parsimonious with each element following exponential AR(1) Gaussian process. The number of parameters to be estimated is of order $2m + m^2$. If the off diagonal elements of Σ_n are assumed to be zero, the model reduces to a Constant Conditional Correlation model (Bollerslev,1990). Otherwise the elements of h_t are dependent.

2.4. Volatility-Forecasting Using High-Frequency Data and Implied Volatility

Additional to the traditional models used to capture volatility movements (GARCH-type or stochastic volatility models), there are other alternatives. These include intra-day data (the sum of squared high frequency returns, see Bingham and Schmidt (2006) for empirical applications) and implied volatility which have also contributed towards a better understanding of future volatility expectations. For the latter method, it is only recently that due to the increased interest in option markets, views regarding asset volatility can be expressed using information implied in the related derivative products. Using however, the option implied volatility to forecast future asset volatility necessitates the assumption of an efficient option market, where volatility remains constant over the entire option's life (Hull and White (1987)). In an earlier study Canian and Figlewski (1993) claim that

implied volatility derived using options on the S&P100 index contain no information regarding expectations of the volatility of the underlying. More recently also try to answer the question whether the S&P500 implied volatility index (VIX) encompasses any further information for future volatility, beyond that offered by model-based forecasts. The answer is not in the affirmative. According to their empirical results, if a considerable large number of models is used, VIX index provides no further incremental information. Blair, Poon and Taylor (2001) reach a different conclusion. By studying options on the same index they find that although high frequency data offer significant information power about future volatility, implied volatility, as this is expressed using the daily observations of the VIX index, is a more preferable choice. This conclusion is also supported by Bali and Weinbaum (2005). Their paper compares the performance of discrete-time GARCH models, the VIX index and the conditional extreme value volatility estimator (EVT). Empirically they conclude that GARCH-type models produce inferior results, in comparison with these generated by the other two alternatives, with EVT being the most dominant of all. Finally VIX offers a less accurate characterisation of realised volatility than EVT and GARCH-type models. Giot and Laurent (2007) using options based on both S&P100 and S&P500, asses the information content of the jump/continuous components of volatility when implied volatility is added as an extra regressor. They find that implied volatility subsumes most relevant volatility information while when a GARCH-type model is encompassed in the regression, implied volatility still is more informative. Additionally, Jiang and Tian (2005) studying options written on the S&P500 index and using a model free implied volatility (based on the earlier work of Britten-Jones and Neuberger (2000)) conclude that their approach offers indeed an informative and superior indicator regarding future realised volatility. Martin, Reidy and Wright (2008) however, in their work find that the model free implied volatility (as this is represented first by Britten-Jones and Neuberger (2000) and afterwards by Jiang and Tian (2005)) is a poor performer. On the contrary, at-the-money options demonstrate superior predictive power of the future volatility for three individual equity indexes. The same does not hold however, when the S&P500 index is under scrutiny. Both model free implied volatility and at-the-money options are weak indicators of the volatility of the underlying equity index. Finally it is supported by relevant literature

that high frequency asset realisations dominate stochastic volatility models, as demonstrated by Andersen, Bollerslev, Diebold and Labys (2003), Deo, Hurvich and Lu (2006), Koopman, Jungbacker and Hol (2005), and Pong, Shackleton, Taylor and Xu (2004), among others. On the contrary, there is empirical evidence that the predictive power of high frequency asset realisations is stronger than implied volatility when using foreign exchange data. This has been demonstrated by Andersen et al. (2001), Andersen, Bollerslev, and Lange (1999), Andersen and Bollerslev (1998), and Martens (2002), Taylor and Xu (1997), Pong, Shackleton, Taylor, and Xu (2004), and Neely (2002), who find that intra-day return variances from the foreign exchange market provide incremental information content beyond that provided by implied volatility forecasts. Regarding stock market data, Andersen et al. (2001), Areal and Taylor (2002), Martens (2002), and Fleming, Kirby, and Ostdiek (2003) reach a similar conclusion regarding the superiority of the information power offered by intra-day data. Finally Martens and Zein (2004), studying three different asset classes, equity, foreign exchange and commodities conclude that historical intra-day returns can compete equally and in many instances over-perform implied volatility forecasting ability of the future volatility movements.

What is apparent from the above is that implied volatility suffers from microstructure noise, a problem various authors have tried to account for. The most popular approach is sampling sparsely data as it has been suggested by Andersen and Bollerslev (1998), Adersen, Bollerslev, Diebold and Labys (2000, 2001), Barndorff-Nielsen and Shephard (2001) etc. Using this method however, produces a discretisation error (Barndorff-Nielsen and Shephard (2002a), Meddahi (2002) etc). Improvements of this estimator have been suggested from many researchers, including the first-order autocorrelation adjustment by Zhou (1996), the optimal sampling frequency by Bandi and Russell (2006, 2008) and Ait-Sahalia, Mykland and Zhang (2005), the average and two-scale estimator of Zhang, Mykland and Ait-Sahalia (2005), the multi-scale estimator of Zhang (2006), the realized kernel estimator of Barndorff-Nielsen et al. (2008) etc.

2.5. Estimation Methods

2.5.1. Markov Chain Monte Carlo

A Markov chain is a process of random draws from a distribution such that a future outcome given the present one is independent of any past draw. Markov Chains are aperiodic meaning that we can not identify any parts of the state space that the process takes particular values at specific times, irreducible so that there is unlimited communication and interaction between the states and finally positive recurrent. All above mentioned properties endow the chain with a "unique stationary distribution" which is also the limiting distribution of the chain and the Law of Large Number together with Central Limit Theorem hold.[1] Markov Chains also stimulate the notion of convergence after a sequence of iterations is performed which is not required when importance sampling or other similar methods (weighted bootstrap, rejection sampling) are used. Markov Chain Monte Carlo (MCMC) was initially introduced by Tanner and Wong (1987) to simulate missing values of a process (Data Augmentation algorithm). More generally a MCMC algorithm is a process that is used to simulate distributions through sampling from a Markov Chain that has as its unique stationary distribution the probability distribution of interest. Apparently the main advantage of the method is that sampling is not performed directly from the underlying distribution but from the Markov Chain which is constructed with the help of a transition kernel. The most popular MCMC procedures include the Metropolis-Hastings (Random Walk MH, Independence MH etc.) and Gibbs Sampling algorithms (a special case of the first but not applicable in particular instances).

Suppose for example there are $x = x_1, x_2, \ldots, x_n$ observations with probability distribution $p(x)$ and a proposal distribution $q(y|x)$ while $y = y_1, y_2, \ldots, y_n$ random draws. In the Metropolis-Hastings algorithm we accept the proposal draw with a particular probability,

$$\begin{cases} x_{t+1} = y_t & \text{Prob} = \alpha \\ x_{t+1} = x_t & \text{Prob} = 1 - \alpha \end{cases} \tag{2.153}$$

[1] For reference regarding Markov Chains see Mira (2004)

where the probability α equals,

$$\alpha(\mathrm{x_t}, \mathrm{y_t}) = \min\left\{\frac{\mathrm{p(y)q(y,x)}}{\mathrm{p(x)q(x,y)}}, 1\right\} \tag{2.154}$$

This is introduced by Metropolis *et al.* (1953) but requires the proposal distribution to be symmetric, $q(x|y) = q(y|x)$. Later Hastings (1970) introduces the extended version that relaxes the above mentioned restriction. Some criticism regarding the algorithm is made due to the fact that the proposal distribution is usually carefully chosen so as to be easily simulated while a tuning parameter is also necessary. This tuning differs from one problem to another so the researcher has to make relevant each time adjustments.

The Metropolis-Hastings (MH) algorithm can be either Independent or Random Walk. In the Independent MH the current draw does not depend on previous random draw while a common choice for the proposal distribution is the multivariate Gaussian,

$$\begin{aligned} q(x, y) &= q(y), \\ y &\sim N(\mu, cV), \end{aligned} \tag{2.155}$$

where $\mu = \tilde{\theta} = \arg\max \ln p(\tilde{\theta}|y)$, c is the tuning parameter and $V = -c\left\{\left[\frac{\partial^2 \ln p(\theta|y)}{\partial\theta\partial\theta'}\right]_{\theta=\mu}\right\}^{-1}$, and retained with probability (see Geweke (1994), Amisano and Geweke (2006)),

$$min\left\{\frac{\mathrm{p(y)p}(x|\mathrm{y})\mathrm{q}(\mathrm{y_{t-1}},\mu,\mathrm{cV})}{\mathrm{p}(\mathrm{y_{t-1}})\mathrm{p}(x|\mathrm{y_{t-1}})\mathrm{q(y},\mu,\mathrm{cV})}, 1\right\} \tag{2.156}$$

where μ is a vector with model's parameters ($\mu = \tilde{\theta}$).

The Random Walk MH instead takes under consideration only symmetric proposal distributions,

$$q(x|y) = q(y|x). \tag{2.157}$$

With no restrictions for the variance covariance matrix it retains draws with probability

$$min\left\{\frac{\mathrm{p(y)p}(x|\mathrm{y})}{\mathrm{p}(\mathrm{y_{t-1}})\mathrm{p}(x|\mathrm{y_{t-1}})},1\right\}. \tag{2.158}$$

Although the Independent MH algorithm occasionally yields superior results, the rate of acceptance is higher than that of the Random Walk MH.

2.5.2. Bayesian Inference

Bayes theory involves the specification not only of a model generating the data along with a set of unknown parameters, but also a prior distribution for these parameters.

Let $y = y_1,\ y_2,\ \ldots,\ y_n$ be a vector of independent identically distributed observation with θ unknown parameters. Bayes theorem states that inferences for θ can be made based on the posterior distribution which equals (Carlin *et al.* (2000))

$$\begin{aligned} p(\theta|y,\eta) &= \frac{p(y,\theta|\eta)}{p(y|\eta)} = \frac{p(y,\theta|\eta)}{\int p(y,\theta|\eta)d\vartheta} \\ &= \frac{p(y|\theta)p(\theta|\eta)}{\int p(y|\theta)p(\theta|\eta)d\vartheta} = \frac{l(y|\theta)p(\theta)}{m(y|\eta)}, \end{aligned} \tag{2.159}$$

where $p\,(y|\theta)$ is the probability distribution of y, $p\,(\theta|\eta)$ the prior distribution of the parameters with η hyperparameters. $m\,(y|\eta)$ is also called the marginal distribution and it is not dependent on θ but on η instead. A challenge to surmount, however, lies ahead, regarding the specification of the prior distributions or else the prior belief of the researcher about each component of θ. Usually, frequentists condemn Bayesian enthusiasts due to this dependence on subjective prior beliefs and convenient priors (conjugate). However, it is common knowledge today that the selection of non-informative priors such as uniform can ease the problem if no further details about the parameters' distributional pattern are available. In fact, however, there exists unconstrained choice of any prior distribution due to the computational ease available nowadays. This fact can greatly contribute towards the objectivity of the research.

Apart from parameter estimation, another popular application of Bayes theory is when performing model averaging. The technique, also known as Bayesian Model Averaging (BMA), can easily provide a way to contend

with the uncertainty involved in the model selection procedure.[2] Suppose for example that there are n models to be considered, M_1, M_2, ..., M_n, all of which have a probability density and a likelihood function. The researcher can assign a set of prior probabilities to each one of these models under consideration, Prob(M_i), along with a prior for the parameters θ_i. Here it is assumed that a collection of data points is observed $y = y_t$ (where y represents the returns vector) and the quantity of interest is Δ; then its posterior equals

$$\Pr(\Delta|\mathrm{y}) = \sum_{\mathrm{i}=1}^{2^{\mathrm{n}}} \Pr(\Delta|\mathrm{M_i}, \mathrm{y}) \Pr(\mathrm{M_i}|\mathrm{y}), \tag{2.160}$$

with 2^n candidate models to have generated Δ (n can, as we will see in subsequent chapters, equal the number of different predictors). Using the Bayes rule the posterior for model M_i is as follows:

$$\Pr(\mathrm{M_i}|\mathrm{y}) = \frac{\Pr(\mathrm{y}|\mathrm{M_i})\Pr(\mathrm{M_i})}{\sum_{\mathrm{i}=1}^{2^{\mathrm{n}}} \Pr(\mathrm{y}|\mathrm{M_i})\Pr(\mathrm{M_i})}, \tag{2.161}$$

where $\Pr(\mathrm{y}|\mathrm{M_i}) = \int \Pr(\mathrm{y}|\theta_{\mathrm{i}}, \mathrm{M_i}) \Pr(\theta_{\mathrm{i}}|\mathrm{M_i}) d\theta_i$ is the marginal likelihood of the model M_i and $\Pr(\mathrm{M_i})$ is the prior probability of the model (see Madigan and Raftery (1994), Raftery, Madigan and Hoeting (1999), Tang (2003)). Then the Bayes factor (BF) is defined as (Wasserman (1997))

$$\mathrm{B_{ij}} = \frac{\Pr(\mathrm{M_i}|y, n)}{\Pr(\mathrm{M_j}|y, n)} = \frac{m_i}{m_j} = \frac{\int L_i(\theta_i) p_i(\theta_i) d\theta_i}{\int L_j(\theta_j) p_j(\theta_j) d\theta_j}, \tag{2.162}$$

which can give a coherent measure when comparing model M_i against model M_j (Bayes factors are important for averaging purposes). According to Jeffrey's (1961), the Bayes Factors can be interpreted accordingly, for example $BF > 10$ indicates an apparent dominance of the first model against the second etc. If we had an almost infinite collection of models, averaging all of them could yield better results. However, there are practical problems when trying to implement the above suggestion, which implies that the examiner has to restrict himself in the selection of models only up to a point that is

[2] An important reference for the researcher is the Bayesian Model Averaging homepage (Volinsky (2010)) where one can identify some prominent papers and pieces of literature towards a better understanding of the significance and the evolution of the method.

computationally plausible. Madigan and Raftery (1994) propose a graphical model selection algorithm, which is later extended by Volinsky, Madigan, Raftery (1996) to a linear regression scheme named "Occam's Window" method, according to which models that underpredict in comparison with the best predicting model should not be considered. They also exclude models that are too complicated for the researcher and do not add any further knowledge in comparison to their simpler equivalents. Summarizing, they defined two classes of models:

$$A' = \left\{ M_i : \frac{\max_j \{pr(M_j|y)\}}{pr(M_i|y)} \leq c \right\}, \tag{2.163}$$

where M_i is each individual model and c defines the boundaries that should be surpassed, and

$$B = \left\{ M_i : \exists M_j \in A, M_j \subset M_i, \frac{pr(M_j|y)}{pr(M_i|y)} > 1 \right\}. \tag{2.164}$$

Then the models not belonging in A' and those included in B are left out (see Madigan and Raftery (1994), Volinsky, Madigan, Raftery, Kronmal (1997)), so that the researcher is left with A as the targeted subspace, where,

$$A = A' \backslash B. \tag{2.165}$$

The posterior probability (eq. 2.159) is transformed as follows:

$$\Pr(\Delta \,|y) = \sum_{M_i \in A} \Pr(\Delta \,|M_i, y) \Pr(M_i \,|y). \tag{2.166}$$

The second principle underpinning Occam's Window relates to the posterior odds ration. If model M_0 is smaller than M_1 and there is supportive evidence for M_0, then M_0 is accepted. To reject M_0 and accept M_1 stronger evidence would be required, while when evidence is inconclusive we reject neither of them.

Another alternative to approximate (2.159), suggested by Madigan, York (1995), is using the Markov Chain Monte Carlo Model Composition MC^3 algorithm. Under this method, if we simulate a Markov Chain $M(t)$, with $t = 1, 2, \ldots, N$ representing the random draws, with state space M and equilibrium distribution $\text{Prob}(M_i \,|y)$, then for any function $g(M_i)$ as $N \rightarrow$

∞ the summation $\frac{1}{N}\sum_{t=1}^{N} g(M_t)$ is an estimate of the expectation of $g(M)$.

$$\begin{aligned} \hat{G} &= \frac{1}{N}\sum_{t=1}^{N} g(M(t)), \\ \hat{G} &\rightarrow E(g(M)) \; as \; N \rightarrow \infty. \end{aligned} \tag{2.167}$$

Here in particular the function $g(M)$ equals

$$g(M) = \Pr\left(\Delta \,|M, y\right). \tag{2.168}$$

The method uses the entire model sample space, while we accept a draw from the proposal "neighborhood" state which includes models with one more and one less parameter than the current model M with probability

$$\min\left\{1, \frac{Pr\left(M'\,|y\right)}{Pr\left(M\,|y\right)}\right\}, \tag{2.169}$$

where M' is the proposal state and $M \in \Omega$ where M is the model space.[3]

When referring to the traditional approach of Bayesian Model Averaging where Bayes Factors are involved, the calculation of the marginal likelihood is always an arduous task to perform. The integral representing the marginal distribution is only numerically tractable in the majority of the cases. Instead of estimating the Bayes Factors, however, we could approximate the method. In the past a variety of alternatives have been proposed which approximate BMA. Some of these include the Laplace Method (Tierney and Kadane (1986)), BIC approximation (Schwarz (1978), Madigan and Raftery (1995), Kass and Wasserman (1995)) and finally the MLE approximation (Draper (1995), Raftery *et al.* (1996), Volonsky *et al.* (1997)).[2] There are however cases where these approximations fail to provide a competitive result; thus the calculation of the actual marginal likelihood is again under consideration in order to perform Bayes inference. A convenient way to calculate the posterior density is through the use of MCMC algorithm. Chib (1995) proposes an approach when MCMC with Gibbs Sampling is used which is extended (2001) to include MCMC with the Metropolis-Hastings

[3] A more detailed analysis can be found in Hoeting *et al.* (1999), Madigan, & Raftery (1994), Madigan,& York (1995) and Volinsky *et al.* (1997) while for the Metropolis-Hastings algorithm see Chib,& Greenberg (1995).

algorithm. Gelfand and Day (1994) also offer an approximation of the integral that can be applied in both cases, while Geweke (1999) modifies their early work by suggesting a truncated multivariate distribution to be used when calculating the marginal distribution.

3. Volatility Model Averaging and Option Pricing

3.1. Introduction

For over three decades conditional volatility forecasting has undeniably attracted a considerable amount of interest. However, there exist significant differences in prediction outcomes, since they greatly depend on the particular model choice. Evidence of the existence of a dominant proposal cannot easily be traced, while, the ever changing market conditions increase the generic problem with volatility forecasting. Model uncertainty can cause direct complications, not only to option pricing as it will be discussed here, but also to a vast number of financial applications including risk management related decisions and asset allocation problems. The idea of averaging can be traced back to 1963 when Barnard observed that by simply averaging two separate models he could enhanced forecastability. Additionally, Bayesian Model Averaging (BMA) can be attributed to Leamer (1978), whose work is a continuation of Robert's (1963), and suggests the use of combined model posterior distributions. A more profligate future for Bayesian enthusiasts under the discrete-time volatility spectrum could not be foreseen before the early 90s, when Geweke (1994) used importance sampling in order to implement Bayesian Inference for a GARCH model. Although Bayesian Model Averaging is promising, it makes assumptions of the prior and proposal distribution of the model's parameters and imposes to the researcher an unavoidable computational burden. In order to ease calculations, an approximation of the above method is usually used, where the Bayes Factor is substituted by information criteria, when the method is referred to as Bayesian Approximation (BA). In a more contemporary work, Granger and Jeon (2003) propose the use of Thick Modelling (TM) to perform averaging.

This chapter first presents the three mentioned distinct approaches of averaging (BMA, BA, TM) and how they perform compared to their single model alternatives. We also propose two new methods to determine the weights assigned to each model for Thick Modelling, first using an optimisation algorithm and second using Neural Networks. These approaches try to capture any nonlinear relationships between the various methods used for forecasting. The number of competing volatility models is more than 180, all used in a ground-breaking attempt to attain forecasting optimality. This is a daunting task to perform, especially when Bayesian Model Averaging is under consideration. Results though are promising. It is also worth mentioning that Bayesian inference has been applied only to the simple GARCH(1,1) model.[1] This created an additional impediment when averaging a wide variety of complicated discrete volatility proposals under the Bayes theory spectrum.

Accurate volatility forecasting is also fraught with significance when option pricing is under consideration. Using as reference the Heston & Nandi (2000) model for discrete volatility option pricing, two further novelties are suggested: first pricing *by averaging the individual option prices under different assumptions of the underlying volatility*, and second, pricing *by averaging* not the set of all prices this time but *the alternative volatilities that drive the dynamics of the underlying asset, and ultimately feed this averaged volatility to the model and then price.* Both proposals are contrasted against single volatility pricing models. What is also novel here is the fact that every previous attempt regarding pricing options with discrete volatility assumption of the underlying has been focused only on simple models from the GARCH family (Duan (1995, 2000), Christofersen (2004)), and only up to lag one. This narrow view is extended here so that some of the most important conditional-volatility models are estimated, extending their dimension up to lag four.

The rest of the chapter is developed as follows. First some literature regarding the application of Bayes theory when estimating conditional-volatility is presented. Afterwards all different methods of averaging used in the research, including Bayesian Model Averaging (BMA), the Bayesian

[1] The only exception is Vrontos *et al.* (2000), who also present EGARCH but they restrict the analysis to one lag for the GARCH model and to the second lag for the EGARCH.

Approximation (BA) and a considerable number of Thick Model Averaging (TMA or TM) alternatives, are presented. Comparison of these averaged schemes with single volatility models follow, while further details regarding the theory underlying their implementation are provided. In the second part, option pricing becomes the focal point. Here we compare the model constructed by averaging processes with different assumptions regarding the volatility dynamic of the underlying with the one that has the averaged volatility alone as its driving force, as well as with single model volatility proposals. An empirical application that involves forecasting the volatility of the S&P500 index, as well as pricing options written on the same index, substantiates the superiority of the averaging propositions against any single alternative.

3.2. Literature Review

3.2.1. Bayesian Model estimation & averaging for conditional-volatility models

One of the most important applications of Bayes theory is when modelling conditional heteroscedasticity. First Geweke (1989, 1994) applied the importance-sampling method, and later the Metropolis-Hastings algorithm, in order to carry out inference on GARCH models. Bauwens and Lubrano (1996), based on some inherent drawbacks of the Metropolis-Hastings algorithm, tried to use the Griddy-Gibbs sampler (Ritter and Tanner (1992)) so as to apply the method for the GARCH (1,1) model with Student-t innovations and account for fat tails and excess kurtosis. They also compared their results with those acquired when performing Importance Sampling and Metropolis-Hastings algorithm instead. Finally, they provided some applications using an asymmetric GARCH on the Brussels' stock exchange index and predicted option prices. Nakatsuma (1998, 2000) applied a Markov-Chain sampling technique with the Metropolis-Hastings algorithm in order to perform Bayesian estimation and analysis (non-stationarity test etc.) for GARCH models. Numerical applications of the model include simulated data as well as weekly foreign exchange rates of the US dollar to Swiss Franc. At the same time, Vrodos, Dellaportas and Politis (2000) moved a step forward by estimating GARCH and EGARCH models and

used their posterior average distribution to forecast volatility. They applied the reversible-jump MCMC for the implementation of the Bayesian inference, and they concluded their paper with applications related to the Athens stock exchange index and to option pricing. Following Bauwens and Lubrano (1996), Concepcion-Ausin and Galeano (2006) used Griddy-Gibbs sampler to estimate the GARCH model but with innovations that follow a Gaussian mixture distribution. According to the authors, although there are some problems when analyzing mixture models, Bayes theory can facilitate and ease the estimation of the likelihood function for the parameters and consequently the evaluation of the posterior distribution. The Swiss Market Index is used for application of the Gaussian mixture GARCH(1,1) model and some problems related to the estimation of the VaR are also addressed. Finally Verschuere (2007) introduced a MCMC algorithm in order to estimate Markov-Switching GARCH models (Hamilton and Susmel (1994), Gray (1996)). The particular method can be viewed as a regime-switching model that allows correlation between the regimes. Here the Bayesian approach facilitates the estimation of this extra parameter that governs volatility regimes using Gibbs sampler algorithm (see Smith and Roberts (1993) for Bayesian estimation of hidden Markov Chains). The data used to test the model are simulated from a Monte Carlo process. However, suggestions for further implementation of the method using real data are made, along with its comparison with the one presented in Gray's paper (1996) for regime switching GARCH models.

3.2.2. Thick Model Averaging for conditional-volatility models

In contrast with Bayesian Model Averaging, Thick Modelling has only been introduced recently by Granger and Jeon (2003). What has been suggested initially is the use of a simple average of the top (according to a selected percentile) performing models. The intuition behind their suggestion is the fact that there might be a model that is not optimal in a particular time period but might under different circumstances. Considering the highly volatile market environment it comes as a natural conclusion that a catholic dominant forecasting method will always be elusive, so instead one can opt for averaging the strongest canditates.

Initially, let f_t^i be the forecast of model i at time t where there exist n different modelling alternatives competing to be selected from the researcher. Thick Model Averaging states that instead of opting for a single model one can simply average all of them and based on that averaged scheme produce the final forecast:

$$F_t = \frac{1}{n}\sum_{i=1}^{n} f_{i,t}. \tag{3.1}$$

An alternative to the above simple approach is to choose a percentage of the best and worse performing models that you want to be removed in an attempt to eliminate any extreme performances. Trimming is something proposed by Stock and Watson (1999) and can be expressed as follows,

$$F_t = \frac{1}{n - Trim}\sum_{i=1}^{n-Trim} f_{i,t}. \tag{3.2}$$

$Trim$ is the number of forecasts that were trimmed. In order to assign more weights to the better forecasts and less to the worse, OLS has also been suggested instead of simple averaging. There are however multicollinearity problems reported under this approach. A possible solution to that is to exclude some of models from the regression equation and repeat calculations. In order to find the combined forecasts there are two steps involved. First we estimate the parameters of the model by performing the regression:

$$(r_t - \mu)^2 = \sum_{i=1}^{n} \hat{w}_i f_{i,t} + \varepsilon_t. \tag{3.3}$$

Afterwards we can combine the weighted forecasts according to the estimated coefficients and produce the final averaged prediction, which now equals

$$F_t = \sum_{i=1}^{n} \hat{w}_i f_{i,t}. \tag{3.4}$$

Another approach to find the weights is by using a loss functions. The choice of the loss measure is subjective, though. Let l be a loss function; then the weights can be estimated as,

$$\hat{w}_i = \frac{l_{i,t}}{\sum_{i=1}^{n} l_{i,t}}. \tag{3.5}$$

3.2.3. Neural Networks and conditional-volatility forecasting

The ability to forecast volatility using neural networks has long been a controversial issue. Although artificial intelligence related techniques for forecasting were first introduced in the late 80s-early 90s, they were regarded by many as obscure or else a "black box". No less than a decade ago, neural networks have been back in the spotlight, regaining the lost territory in our collective consciousness as an efficient technique for predicting the future. This is mainly driven by the recognition of their ability to capture adequately complex nonlinearities that underly many time series and ultimately to outperform their linear rivals.

Despite the various Artificial Neural Networks (or else Neural Network models) that can be found in the literature, the answer to the question which one of those performs better still is not straightforward. The most commonly used, however, is the Artificial Neural Network with Feed Forward Back Propagation algorithm (FFBP). Its efficiency has been a subject of research more than for any other Neural Network model. By some recent estimates it accounts for the 90 percent of all Neural Network related applications, while in particular occasions it performed adequately well. The development of other alternatives including Hopfield networks (Recurrent), Probabilistic, Fuzzy Logic, Self Organized Networks etc. also share a part in the endeavor to make more accurate predictions under the Neural Network scope.

Early work on volatility forecasting using Neural Networks can be found in Catfolis (1996). What is attempted is the prediction of stock return volatility implementing a Nonlinear AutoRegressive Moving Average with eXogenous inputs model (NARMAX) with Neural Network structure. The results for a forecasting horizon of four weeks were satisfactory; however, the author points out that different structures have to be implemented in different stocks. Thus a generalization of a catholic architecture to the whole market is never optimal. In the same year Ormoneit and Neuneier (1996) used neural networks to predict the volatility of the German stock index DAX using first multilayer perceptrons and second density estimating NN. The results indicated predictive dominance of the second method, as density-related NN can model efficiently complicated residual probability models. Harrald and Kamstra (1997) moved one step forward by evolving Artificial NN to combine volatility forecasts and compared their results with

a naïve average combination, a least squares method and a non-parametric Kernel method. Again their best evolved network (self adaptive instead of simply evolutionary) highly outperformed the other alternatives. Moreover their suggestion exhibited supremacy when statistical tests for encompassing were performed. In the same year Freisleben and Ripper introduced a new model that in addition to predicting the conditional expectation could also 'learn' the conditional variance. To asses its performance a comparison with NGARCH is executed and results are once more satisfactory. Additionally, Schittenkopf and Dorffner (1998) used mixture density networks to predict volatility, incorporating the long-memory characteristic of volatility in these models as well as non-Gaussian probability densities. Results indicated a superiority in comparison with commonly used models for conditional-volatility such as GARCH,GJR etc. More recently Tino, Schittenkopf and Dorffner (2001) were involved with Recurrent NN (RNN) when used to simulate daily straddles, where performance is based on the volatility of the underlying assets. What is demonstrated is that RNN have a restrictive character that prevents the full use of their representation power. This has as a consequence that they either overestimate noisy data or have a finite low memory. Thus they conclude that there is no apparent reason against using simpler fixed-order Markov models instead. Additionally, Hamid and Iqbal (2002) have written a primer for NN when used to forecast time-dependent volatility. They use a simple FFBB NN which significantly outperformed implied volatility and is broadly consistent with realized volatility. The basic intuition behind this paper is the demonstration that even a basic structure of a NN is adequate to offer forecasting dominance. Finally Roh (2006) managed to enhance the predictive power of classical financial time series models as EWMA GARCH and EGARCH by integrating them into an Artificial NN. Additionally the author employed a two-sided focal point regarding not only the deviation but also the direction of the forecasted variables. Finally in this study after the coefficients were estimated using conventional financial time series models, new variables that affect the results are extracted. This is in contradiction with classical techniques that use trial and error to find the optimal coefficients of the inputs in order to produce the best forecasting output.

3.3. Methodology

3.3.1. Methodology for Bayesian Volatility Estimation

From what has already been mentioned regarding the literature, any attempt that has been made so far is strictly restricted in the use of simplified models for estimation convenience. The scope of this research, however, extends beyond model and lag confinements by using Bayesian Inference to estimate a wide variety of the most important GARCH models and for every specification of them (every possible combination up to lag(p,q) four). An additional contribution however is the fact that Bayes theory is also used to account for model uncertainty by applying Bayesian Model Averaging. This endeavor generates a collection of more than 180 single models to estimate and to average them using their posterior probabilities. Since there are only few papers that implement Bayesian Model Averaging and none in the field of volatility forecasting, this is a novice attempt that contributes to a better understanding of the use of the theory in attaining superior forecastability of conditional-volatility. Until now the frequentist approach has almost monopolistically governed the bulk of the research, due to ignorance related to the drawbacks of the method as well as for its computational ease. Contemporary technological advances, however, emancipate the researcher who is capable of reaping the benefits of hi-tech expertise and yield superior forecasts under the Bayes spectrum. However, to account for some unavoidable discrepancies, BMA is also contrasted with other competing averaging alternatives, including Bayesian Approximation and Thick modelling (Granger and Jeon (2003)).

In particular, the volatility models that are investigated here include:

- GARCH(p,q) $\forall p,\ q = 1, \ldots, 4$;
- APGARCH(p,q) $\forall p,\ q = 1, \ldots, 4$;
- AVGARCH(p,q) $\forall p,\ q = 1, \ldots, 4$;
- NARCH(p,q) $\forall p,\ q = 1, \ldots, 4$;
- NAGARCH(p,q) $\forall p,\ q = 1, \ldots, 4$;
- TGARCH(p,q) $\forall p,\ q = 1, \ldots, 4$;

- GJR-GARCH(p,q) $\forall p,\ q = 1, \ldots, 4$;
- EGARCH(p,q) $\forall p,\ q = 1, \ldots, 4$;
- Fattailed GARCH(p,q) $\forall p,\ q = 1, \ldots, 4$;
- FIGARCH(p,d,m) $\forall p,\ q = 1, \ldots, 4$;

The last model accounts for the phenomenon of long memory (persistence) when attempting volatility forecasting. To ensure positivity and stationarity of the variance, we have to restrict the models' parameters (for further details refer to the original papers). For this reason unbounded support is achieved here by reparametrisation following Amisano (2006):

$$\begin{aligned} \theta = [\ln(\omega),\ & \frac{\ln(\alpha_1)}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{i=1}^{q} \beta_i}, \ \ldots, \ \frac{\ln(\alpha_i)}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{i=1}^{q} \beta_i}, \\ & \frac{\ln(\beta_1)}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{i=1}^{q} \beta_i}, \ \ldots, \ \frac{\ln(\beta_j)}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{i=1}^{q} \beta_i}] \\ \forall i = 1, \ \ldots p \text{ and } \forall j = 1, \ldots, \ q \text{ where } & max(p, q) = 4 \end{aligned} \tag{3.6}$$

The next step in order to begin applying Bayesian Inference is to make distributional assumptions for the priors of the parameters. Prior distributions are in general a subjective choice, although uninformative priors (Uniform) can establish a more unbiased scope of the research. Empirical evidence, however, favours the use of the Gaussian distribution, which can be asymptotically a sufficient assumption even when it is not the true underlying distribution. Gaussian priors are conjugate to themselves, dictating the posterior to be Gaussian as well.

In general, according to Bayes Theory the posterior distribution of θ conditional on past realizations equals

$$p(\theta|y, \eta) = \frac{l(y|\theta)p(\theta)}{m(y|\eta)}, \tag{3.7}$$

where $l(y|\theta)$ is the likelihood function and $p(\theta)$ is the prior density of the parameters with η hyperparameters. In most cases, however, the analytical estimation of the posterior distribution is not possible, so alternative approaches have to be adopted. Here a Markov Chain Monte Carlo algorithm

is used to facilitate the calculations by sampling from the proposal posterior draws for the parameters. The posterior mean can be regarded as an estimate of the parameters $\theta_1, \theta_2, \ldots, \theta_i$, where the length of i depends on the model. More generally,

$$\theta = E(\theta|y, \eta) = \frac{\int \theta p(\theta) l(y|\theta) d\theta}{\int p(\theta) l(y|\theta) d\theta} = \frac{1}{T} \sum_{t=1}^{T} x_t, \tag{3.8}$$

where T is the number of draws from the proposal and x_t symbolizes the random number (draw). The main characteristic of a Markov Chain is that the next draw depends only on the current draw, autonomously of any past draws, and is retained with a particular probability α or else it is dependent on a transition kernel. As the number of draws increases the chain will converge eventually to a distribution which is the "true" posterior distribution of interest. [2]

From the above discussion there is, as expected, a plethora of issues that dictates our attention to be drawn at. First the *selection of prior* is usually favoured to be Gaussian or Student-t. This is mainly due to the fact that these are convenient computationally distributions, not always preserving the objectivity of the method. Here the Normality assumption of the prior distribution of parameters is considered reasonable, since the number of draws is sufficiently large.[3]

Afterwards, the *choice of the starting values for the parameters* should be put under scrutiny. Since the chain after a considerable large number of iterations will converge to its unique stationary distribution, the particular starting values of the parameters tend to play a trivial role towards the fulfillment of this objective. Gilks *et al.* (1996) mentioned that "a rapidly-mixing chain will quickly find its way from extreme starting values. Starting values may need to be chosen more carefully for slow-mixing chains to avoid a lengthy burn in". However it is common practice to start with the estimates derived using maximum likelihood to facilitate quicker convergence.

The burn-in sample that has been mentioned previously refers to the number of iterations that are excluded before averaging the random draws to

[2] See Figure 3.2, 3.3 for the convergence plots of parameters to their theoretical distribution using MCMC.

[3] For further discussion about noninformative priors or improper priors refer to Jeffreys (1961), Kass & Wasserman (1995), Wasserman (1997).

take posterior parameter estimates. It refers to the initial part of the chain where abrupt moves are observed that are so extreme that they can influence the final result, since values from there can act as outliers. The *size of the burn-in-sample* usually depends on the starting values of the parameter estimates, which also determine how quickly the algorithm converges to the desired posterior distribution. Another factor to be taken into account is how close the proposal distribution is to the posterior one of interest. In practice, however, there are a lot of assumptions to be taken under consideration; thus the determination of the length of the burn-in-sample is far from trivial to calculate. The most popular and easy to apply method is to observe when the chain becomes stable enough and discharge the part of the chain that is most volatile (usually, and what we also use here, is around the 10% of the observations to be discharged, though others suggest a smaller percentage). Another way is to use convergence diagnostics (Cowles and Carlin (1994), Brooks and Gelman (1998)). Finally the use of multiple parallel chains until they become close can provide a way to access the size of the burn in-sample. Conclusively the determination of the size of the sample to be excluded as the chain approaches stationarity is so important that if the number of simulations is not sufficiently large and the relative burn in is relatively small then incorrect predictions of the parameters can be produced.

After the above considerations, the Random Walk Metropolis-Hastings algorithm is selected in order to facilitate the simulation of the posterior. The scaling (or tuning) parameter for this algorithm plays a role of immense importance. For example, a miscellaneous choice can either produce high acceptance rates where small steps dictate only a small part of the distribution to be visited, or low acceptance rates where the algorithm moves from the centre of the distribution to the tails abruptly, leaving a lot of space not adequately mapped.

Additionally the update of the parameter vector is performed simultaneously and not one at a time (Single Component Metropolis algorithm, Metropolis *et al.* (1953)), according to Vrontos *et al.* (1999), who proved that it is more efficient to sample simultaneously from the full proposal for each component of the parameter vector. The reason why at first place the use of a Single Component update is advocated is due to the fact that the subdivision of θ into parts is thought to facilitate the reduction of the

computational burden. However, this approach ignores the fact that the components of the θ vector are highly correlated, which when disregarded can significantly decrease the performance of the algorithm (see also Hills and Smith (1992)). [4]

3.3.2. Methodology for Bayesian Model Averaging

Undeniably BMA has been a popular choice as a method to alleviate the uncertainty of selecting one particular model to describe the observing data and to yield accurate forecasts based on this preference. Usually, when the object of interest is volatility there exists an abundance of models which can be used in the averaging process. If we assume for example that there are two competing choices, M_1 and M_2 where $M_1 \neq M_2$, then we can use the following ratio to make relevant comparisons:

$$\frac{\Pr(M_1|X)}{\Pr(M_2|X)} = \frac{\Pr(M_1)\int p(y_1)p(y_1|X)dy_1}{\Pr(M_2)\int p(y_2)p(y_2|X)dy_2} = \frac{\Pr(M_1)m(X|M_1)}{\Pr(M_2)m(X|M_2)}. \tag{3.9}$$

The first part of the fraction, $\frac{\Pr(M_1)}{\Pr(M_2)}$, is called the "prior odds ratio", while the second part, $\frac{m(X|M_1)}{m(X|M_2)}$, is the "Bayes factor" (Gelfand and Day (1994), Chib (1995, 2001), Geweke (1996)). This "posterior odds ratio" can be estimated for all n competing models in order to help us determine the most dominant one. Since the prior odds ratio is assumed to be the same for every pair (in order to guarantee subjectivity), what is left to the researcher is the calculation of the marginal likelihoods. Among the most important proposals for this reason we can identify Chib's method (1995) which could only be applied to the Gibbs Sampling algorithm while it is later (2001) extended to the Metropolis-Hastings algorithm as well. Another popular method to estimate marginal likelihoods is the one suggested by Gelfand and Dey (1994).

Here we adopt a modified version of what has been originally introduced by Geweke (1996). At first place Gelfand and Dey applied an analytical approximation in order to calculate the marginal distribution. However, the accuracy of these approximation is in doubt when the sample size is

[4] See Figure 3.1 which is an example plot of the actual parameters' distributions against the fitted proposal distributions using MCMC for one of the volatility models studied here (GARCH(1,2)).

small, so they use instead a MCMC approach in order to produce better results. The theory involves the adoption of any function of the parameters vector θ, $g(\theta)$ for example, that can be characterized as a distribution and transform it to give the desired marginal. What is produced indeed if we integrate this selected function is the inverse of the marginal,

$$\begin{aligned}
\int g(\theta)d\theta &= \int g(\theta)\frac{p(\theta|y,\eta)}{p(\theta|y,\eta)}d\theta \\
&= \int g(\theta)\frac{m(y|\eta)}{p(\theta|\eta)p(y|\theta,\eta)}p(\theta|y,\eta)d\theta = 1, \qquad (3.10) \\
m(y|\eta)^{-1} &= \int \frac{g(\theta)}{p(\theta|\eta)p(y|\theta,\eta)}p(\theta|y,\eta)d\theta, \\
m(y|\eta) &= \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{g(\theta_i)}{p(\theta_i|\eta)p(y|\theta_i,\eta)}\right\}^{-1}. \qquad (3.11)
\end{aligned}$$

According to the authors the function $g(\theta)$ could be Normal or Student t. What Geweke (1996) proposed is the so called Modified Harmonic Mean where he defined the distribution $g(\theta_i)$ as follows:

$$g(\theta_i) = p^{-1}(2\pi)^{-\frac{k}{2}}|\hat{\Sigma}_\theta|^{\frac{1}{2}}\exp[-\frac{1}{2}(\theta_i-\hat{\theta})'\hat{\Sigma}_\theta{}^{-1}(\theta_i-\hat{\theta})]I(\theta_i\in\Theta_p), \quad (3.12)$$

where $p\in(0,1)$, $\hat{\theta}$ is the mean and $\hat{\Sigma}_\theta$ is the variance:

$$\begin{aligned}
\hat{\theta} &= \sum_{i=1}^{N} w(\theta_i)\theta_i, \qquad (3.13) \\
\hat{\Sigma}_\theta &= \sum_{i=1}^{N} w(\theta_i)(\theta_i-\hat{\theta})(\theta_i-\hat{\theta})', \qquad (3.14) \\
\hat{\Theta}_p &= \{\theta : (\theta_i-\hat{\theta})'\hat{\Sigma}_\theta(\theta_i-\hat{\theta}) \le q_{\chi^2_n}\}, \qquad (3.15)
\end{aligned}$$

where $q_{\chi^2_n}$ is the q-quantile of a chi-square distribution with n degrees of freedom.

3.3.3. Methodology for Bayesian Approximation for Model Averaging

The estimation of the marginal likelihood involved in the calculation of the Bayes Factor, as mentioned before, is not an easy task to perform. Therefore, in order to carry out Bayesian Model Averaging, approximations are used instead. The first one is based on the Schwarz Information Criterion (SIC) introduced by Schwarz (1978). According to it a model should be selected if it maximizes the following decisive factor:

$$SIC = \log(L_{M_i}) - \frac{1}{2} k_{M_i} \log p,$$

where L_{M_i} is the likelihood function for model i $\forall i = 1, \ldots, n$, k_{M_i} is the number of estimated coefficients and p is the number of independent parameters (observations). This is a modification to the criticized Akaike's Information Criterion (1973), where the above factor is defined as

$$AIC = \log(L_{M_i}) - k_{M_i}.$$

The first Information Criterion is also known as the Bayesian Information Criterion (BIC) (some comments and comparison of those two can be found in Stone (1979) and Hannan and Quinn (1979)). The Bayesian Information Criterion (BIC) commonly is used as an approximation of the log Bayes Factor:

$$\log BF_{ij} \approx L_{M_i} - L_{M_j} + \frac{k_{M_i} - k_{M_j}}{2} \log p. \tag{3.16}$$

As discussed by Smith and Spiegelhalter (1980), the BIC can be used as a global approximation to the Bayes Factor and AIC as a local approximation to the Bayes Factor, using Laplace transformation and assuming a general normal prior density. The same authors introduce BF when there is not clear prior information and improper priors have to be used instead. An alternative to model selection criteria of AIC and BIC is offered by San Martini and Spezzaferri (1983). They opt for the model that has the largest utility function (posterior odds ratio combined with another factor for the linear case). Concluding, they provide an average criterion as well. In 1995 O'Hagan proposed an alternative BF, the Partial BF, meaning that only a

part of the data is used for the calculation of the BF while the rest is considered as training sample. O'Hagan divides the data space into two parts; the first provides information for the priors (and facilitates the need to avoid improper priors) and the second part is used to calculate the BF. The problem of defining the right size of the training sample is mainly considered in the Fractional Bayes Factor model (only a part of the likelihood is used as a training set). The FBF suggests an uncomplicated version of Partial BF, which can offer the researcher ease in calculating and convenience in having a direct connection with the likelihood. Considering conjugacy, Kass and Wasserman (1995) introduced the Unit Information Prior, which has as variance of the prior the inverse of the Fisher Information. The use of a normal distribution as the Unit Information Prior produces an approximation to the BIC criterion of the BF with a smaller error order, $O_p\left(p^{-\frac{1}{2}}\right)$. Yet another model selection criterion, the Intrinsic Bayes Factor, is added by Berger and Pericchi (1996), where an improper prior can be used at the beginning and later on they are updated by using the training sample. In contrast with O'Hagan (1995), they suggest the use of each possible combination of the minimal training sample and then average the result. In addition DiCiccio, Kass and Raftery (1997) review some of the most essential methods of simulating posterior distributions by sampling from proposals with the use of MCMC and other similar procedures. Finally there exists a comparison of BF approaches by Mukhopadhyay, Ghosh and Berger (2005).

The above literature review refers to the use of Information Criteria for model selection purposes and also to their use in order to approximate Bayes Factors. Some conflicting opinions regarding which information criterion is more appropriate for model selection do exist. However, the vast majority of the scientific community recognizes in many cases the advantageous position of the Schwarz Information Criterion, due to its direct connection to Bayes methods. This in turn indicates that it can be helpful when hypothesis testing is under consideration or whenever only noninformative priors are available (BIC is not based on any priors) and model selection is under question. This is the reason why BIC is a more favourable option in this research when model selection is performed. For further details regarding all the above refer to Wasserman (1997) and Raftery (1999).

More precisely, model averaging using BIC can be executed as follows.

First, for each model available we compute its BIC using the formula,

$$BIC_{M_i} = -2\log(L_{M_i}) + k_{M_i}\log p \quad \forall i = 1, \ldots, n, \tag{3.17}$$

where p is the number of data and n the number of models and k_{M_i} the number of parameters of the model M_i. The first part of the equation represents the goodness-of-fit measure, the other half the penalty of adding extra parameters. Then according to Kass and Wasserman (1995) the BIC factor can be used to approximate the marginal likelihood of the model (plus an error term):

$$m(y|\eta, M_i) \approx exp(BIC_{M_i}) + \varepsilon \quad \forall i = 1, \ldots, n, \tag{3.18}$$

where $m(y|\eta, M_i)$ is the approximated marginal distribution of model M_i and ε is the error, which can be reduced with the use of some proper priors. Normally it is considered small enough to be disregarded. The last step is to compute the posterior probabilities by dividing the approximated marginal distribution of each model with the sum of the marginals of all models:

$$BF_{M_{i,t}} \approx \frac{m(y|\eta, M_{i,t})}{\sum_{i=1}^{n} m(y|\eta, M_{i,t})} \quad \forall i = 1, \ldots, n. \tag{3.19}$$

3.3.4. Methodology for Thick Model Averaging

The last method to be used in order to perform volatility model averaging is Thick Modelling, proposed by Granger and Jeon (2003). What they suggest is based on what has already been observed almost forty years ago by Roberts (1965) and extensively discussed thereafter by many researchers. This is fact that the use of a single model may be a suboptimal choice when forecasting in comparison with the use of a combination of many different models. The single model approach is usually referred to as Thin Modelling, while when various equations are combined instead is called Thick Model Averaging. [5]

The main difference between Thin and Thick Modelling is that the second method makes out-of-sample predictions using, not one model (as Thin

[5] A comparison of the Thin and Thick Modelling approaches can also be found in Kilic (2005) where some supplementary ways to rank, select and average are introduced.

Modelling suggests) or the weighted average of all according to BMA or BA, but instead a percentage of the best in-sample models. There are two important issues, however, regarding the implementation of the method that need to be discussed. The first refers to the criterion selected to rank the competing alternatives in-sample, or else to the choice of the measure that dictates which model performs best in-sample. This choice naturally is biased. What has been proposed by Pesaran and Zaffaroni (2006) is the use of Information Criteria as they are widely accepted for model selection purposes. A natural extension to the above suggestion is to use any loss function such as Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error etc. The second consideration regarding implementing TMA refers to the problem of weighting after ranking. Granger *et al.* choose a simple average (equal weights) after removing the 5% or the 10% of the best and worst performing models. On the other hand, Pesaran and Zaffaroni trim out only the poor performing models and average the top 25% and 50% of the remaining.

There is, though, lack of a clear scientific consensus regarding which ranking criterion, what weighting and which percentage of top/worst performing models should be retained/trinmmed out. Here we attempt to give an empirical answer to the above dilemmas by comparing a selection of different TMA methods. Additionally we suggest two further alternatives in order to validate the superiority of the method against single models.

In more detail, the first suggestion refers to the weighting problem. Instead of applying a simple average, we set a target (error) function and try minimize this by changing the weighting applied to the best performing models with the help of an optimisation algorithm. In other words, we treat each individual weight assigned to every model included in the subset of the top performing alternatives as a parameter that has to be estimated through minimization of an error function. We call this approach TMA with Optimization. In particular an unconstrained optimisation is used, the Nedler-Mead Symplex method. This is one of the most classical routines, since it works well for various financial problems. The target of this algorithm here is to minimize the in-sample Mean Squared Error of the forecasted volatility. Finally we have to mention that for the actual volatility realizations a proxy is used, since the process in reality is unobservable.

In steps the method can be described as follows:

- Estimate conditional-volatility based on different modelling assumptions.
- Rank all the models available for the analysis according to the Bayesian Information Criterion.
- Determine what percentage of them will be kept and what we will discharged. Here the retention percentage ranges from $10\%, 20\%, \cdots, 50\%$ for comparative reasons.
- Assign weights to these retained models with the use of an optimisation algorithm which minimizes the in-sample Mean Square Error.
- Combine the forecasts to produce one single averaged volatility prediction,

$$h_{t+1} = \sum_{i=1}^{\tilde{n}} \hat{w}_i h_{i,t+1}, \tag{3.20}$$

  where $\tilde{n} = pn$, with n the total number of models and p the retention percenatage which ranges from 10% to 50%.

The second proposal uses instead of an optimisation algorithm Neural Networks (NN) to facilitate the determination of the optimal weights assigned to each element included in the trimmed model set. The type of Neural Network used in particular is an Artificial Neural Network with Feed Forward Back Propagation algorithm. In order to apply this particular network, first a trial-and-error method is used to verify the number of neurons as well as the number of layers to be used. Here empirical results support evidence in favor of one layer with ten neurons. The activation function employed to propagate the scaled data to the next layer is the tanh sigmoid function (this together with the logistic sigmoid function are widely used due to their convenient derivative). The function is defined as follows:

$$\tanh(y) = \frac{e^y - e^{-y}}{e^y + e^{-y}}. \tag{3.21}$$

Since we have the above activation function, it is natural to assume that the scaling of the data is from -1 to 1. The training of the algorithm is done using Back Propagation (Rumelhart and McClelland (1986)), and it is a supervised training since as output target we set the volatility proxy.

The implementation steps are similar to the ones followed for TMA with Optimization. Here, however, instead of an assumption of linearity for the weights we employ a nonlinear relationship based on the tanh sigmoid function. The intuition behind using NN for averaging is to test and capture any nonlinearities in the weighting possess. Also NN are widely popular for their non-parametric properties, which makes them attractive for higher-dimensional problems.

3.4. Volatility Model Averaging and Option Pricing

3.4.1. Literature Review

The quest to discover the model that accurately describes the elusive process of option prices has led many researchers to deviate from the continuous-time stochastic-volatility modelling regime and embrace discrete-time models. Duan (1995) is the first to propose option pricing where the underling's volatility follows a GARCH process. There are two dynamics that define this model under the physical measure P:

$$\ln S_t = \ln S_{t-1} + r + \lambda\sqrt{h_t} - \frac{1}{2}h_t + \varepsilon_t \quad \forall t = 1, \ldots, n, \tag{3.22}$$

$$h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j} \quad \varepsilon_t \sim N(0, h_t). \tag{3.23}$$

Duan suggests r as the "one-period continuously compounded risk-free rate" and λ as the "constant risk premium". The next step involves transition from the physical probability measure P to the risk-neutral measure Q in order to perform pricing, while the dynamics of the underling now are specified as follows:

$$\ln(S_t) = \ln(S_{t-1}) + r - \frac{1}{2}h_t + \xi_t, \tag{3.24}$$

$$\xi_t|\phi_{t-1} \sim N(0, h_t),$$

$$h_t = \omega + \sum_{i=1}^{p} a_i(\xi_{t-i} - \lambda\sqrt{h_{t-i}})^2 + \sum_{j=1}^{q} \beta_j h_{t-j}. \tag{3.25}$$

In order to reduce computational load, Duan (1997) and later Duan, Gauthier and Simonato (1999), provide an analytical approximation to price European options using the above formulae when the variance is described by a NGARCH process. As mentioned by Duan later, this analytical approximation can be extended in order to include volatilities described by other discrete processes as for example GARCH, EGARCH or GJRGARCH (Duan (2005)). These analytical approximations are formed based on the Black-Scholes (1973) and Merton (1973) pricing equations adjusted for skewness and kurtosis. However, they demand burdensome algebraic calculations.

Following Duan, Heston and Nandi (2000) introduce a closed form solution this time, for European options whose underlying asset has conditional variance that follows a particular GARCH process. The model is based on two major assumptions. The first refers to the log price of the underlying which is described by the following two equations:

$$\ln(S_t) = \ln(S_{t-\Delta}) + r + \lambda h_t + \sqrt{h_t} z_t, \tag{3.26}$$

$$h_t = \beta_0 + \sum_{i=1}^{q} \beta_i h_{t-i\Delta} + \sum_{i=1}^{p} \alpha_i \left(z_{t-i\Delta} - \gamma_i \sqrt{h_{t-i\Delta}} \right)^2, \tag{3.27}$$

where $z_t \sim N(0,1)$ and represents the innovations, r is the risk-free rate and λ is the risk premium, as also defined before in Duan's model. It is worth mentioning that as the time interval Δ shrinks, the conditional variance converges to the continuous-time stochastic-volatility model of Heston (1993). The transition from the subjective to the objective probability measure yields the following transformed dynamics for the underling:

$$\ln(S_t) = \ln(S_{t-\Delta}) + r - \frac{1}{2} h_t + \sqrt{h_t} z_t^*, \tag{3.28}$$

$$h_t = \beta_0 + \sum_{i=1}^{q} \beta_i h_{t-i\Delta} + \sum_{i=1}^{p} \alpha_i \left(z_{t-i\Delta} - \gamma_i \sqrt{h_{t-i\Delta}} \right)^2 + ... + \alpha_1 \left(z_{t-\Delta}^* - \gamma_1^* \sqrt{h_{t-\Delta}} \right)^2, \tag{3.29}$$

where $z_t^* = z_t + \left(\lambda + \frac{1}{2}\right) \sqrt{h_t}$ and $\gamma_1^* = \gamma_1 + \lambda + \frac{1}{2}$. The connection between the previous two equations with these of the risk-neutral measure

becomes apparent in setting $\lambda = -\frac{1}{2}$ and $\gamma_1^* = \gamma_1 + \lambda + \frac{1}{2}$. This is also part of the second assumption made by the authors where z_t^* is assured to be standard normally distributed. This means the option value follows the Black-Scholes-Rubinstein formula. The derivation of the closed form solution shares common ground with Duan's method. Although the model performs sufficiently well empirically, there are strong assumptions underlying the model. These need to be relaxed in order to achieve more catholic applications involving various GARCH-type volatility assumptions.

In this direction, Christoffersen and Jacobs (2004), after categorizing the various specifications of GARCH models according to Ding et al (1993) and Hentschel (1995) and based on the previous work of Heston and Nandi (2000), try to compare different GARCH models for option valuation. There is not, however, a unique analytical solution for all models categories, so these are estimated using Monte Carlo. In an even more contemporary work of Barone-Adesi, Engle and Mancini (2006) a more robust model for option pricing under the GARCH-type volatility spectrum is proposed.

Conclusively the assumption related to the innovation term for our research is relaxed in order to deviate from the usual one of the standard normal, while the GARCH-type model parameters are calibrated using Monte Carlo.

3.4.2. Methodology for Option pricing

Having under consideration all the above mentioned about discrete option pricing methods, it is evident that different models offer different forecasting potential, which is unavoidably dictated by the various model assumptions. Some may adroitly assume incomplete markets where innovations are not standard normal, but fail to highlight why a particular GARCH process should be adopted. Model uncertainty of the conditional-volatility of the underlying has again to be addressed. What indeed provokes the problem is that there is no guarantee for the researcher that for different asset classes, time intervals and frequencies, the same model will always yield an optimal forecast for the option price.

Before performing the volatility averaging, the formulae used to estimate the option prices under different GARCH-type model assumptions for the volatility process of the underlying should be established. The dynamics of

the asset return under the physical measure P are defined as follows:

$$\ln \frac{S_{t+1}}{S_t} = r + \lambda \sqrt{h_{t+1}} - \frac{1}{2} h_{t+1} + \varepsilon_{t+1}, \tag{3.30}$$

where ε_{t+1} is the innovation process with $\varepsilon_{t+1} = \sqrt{h_{t+1}} z_{t+1}$ and $z_t \sim N(0,1)$, r is the continuously compounded risk free rate and λ the risk premium. For the volatility process here various specifications have been assumed including those that follow.

- GARCH

$$h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}. \tag{3.31}$$

- Nonlinear Asymmetric GARCH

$$h_t = \omega + \sum_{i=1}^{p} a_i (\varepsilon_{t-i} - c)^2 + \sum_{j=1}^{q} \beta_j h_{t-j}. \tag{3.32}$$

- Nonlinear GARCH

$$\left(\sqrt[\delta]{h_t} \right)^2 = \omega + \sum_{i=1}^{p} \alpha_i \left| \varepsilon_{t-i} \right|^\delta + \sum_{j=1}^{q} \beta_j \left(\sqrt{h_{t-j}} \right)^\delta. \tag{3.33}$$

- Absolute Power GARCH

$$\begin{aligned} \left(\sqrt[\delta]{h_{t-j}} \right)^2 &= \omega + \sum_{i=1}^{p} a_i \left| \varepsilon_{t-i} \right|^\delta - \gamma \left| \varepsilon_{t-i} \right|^\delta I \left\{ \varepsilon_{t-i} < 0 \right\} + ... \\ &+ \sum_{j=1}^{q} \beta_j \left(\sqrt{h_{t-j}} \right)^\delta. \end{aligned} \tag{3.34}$$

- GJR GARCH

$$h_t = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \gamma \varepsilon_{t-i}^2 I\{\varepsilon_{t-i} < 0\} + \sum_{j=1}^{q} \beta_j h_{t-j}. \tag{3.35}$$

- Threshold GARCH

$$\sqrt{h_t} = \omega + \sum_{i=1}^{p} a_i \left| \varepsilon_{t-i} \right| + \gamma \left| \varepsilon_{t-i} \right| I \left\{ \varepsilon_{t-i} < 0 \right\} + \sum_{j=1}^{q} \beta_j \sqrt{h_{t-j}}. \tag{3.36}$$

- Absolute Value GARCH

$$\sqrt{h_t} = \omega + \sum_{i=1}^{p} a_i \left|\varepsilon_{t-i} - c\right| + \sum_{j=1}^{q} \beta_j \sqrt{h_{t-j}}. \tag{3.37}$$

- Fattailed GARCH

$$h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j} \quad z_t \sim t_n, \tag{3.38}$$

where t_n is the Student-t distribution with n degrees of freedom. The above models incorporate almost every irregularity observed in volatility forecasting, including leverage effects, heavy tails, slowly decreasing autocorrelation coefficients etc. [6]

Following Heston and Nandi (2000), the definition of the risk-neutral dynamics for the asset price is as below:

$$\log\left(\frac{S_t}{S_{t-i}}\right) = r - \frac{1}{2}h_t + \sqrt{h_t} z_t^*, \tag{3.39}$$

where z_t^* is a standard normal random variable, $z_t^* \sim N(0,1)$. The volatility equations have also to be changed, replacing ε_t with ε_t^* where $\varepsilon_t^* = \sqrt{h_t}\left(z_t^* - \lambda\right)$ or $\varepsilon_t^* = \sqrt{h_t}\left(z_t^* - \lambda - c\right)$ for the Nonlinear Asymmetric GARCH and Absolute Value GARCH. The volatility model parameters are transformed as well to represent their risk-neutral equivalents.

The option price, given the above, under the risk-neutral measure equals

$$c(S_T) = e^{-r(T-t)} E^Q\left[\max\left(S_T - K, 0\right)\right], \tag{3.40}$$

$$p(S_T) = e^{-r(T-t)} E^Q\left[\max\left(K - S_T, 0\right)\right]. \tag{3.41}$$

The first equation relates to a European call option, the second one to a European put option. Since there is no analytical solution for the above mentioned formulae, Monte Carlo simulation is used to calibrate the risk-neutral parameters.

The procedure followed for calibration shares common ground with work found in the previously mentioned relevant literature. [7] In more detail,

[6] For more details for each model please refer to the original papers (see references).
[7] Christoffersen and Jacobs (2004), Barone-Adesi, Engle and Mancini (2006).

this is performed in two stages. First the estimation of the GARCH model parameters through maximum likelihood is performed. This is done for each particular model up to lag(p, q) four. These starting values are used as inputs to the optimisation algorithm. The objective is to minimize the in-sample Mean Square Error of the predicted option price when it is compared with the actual price. This is the second step of the procedure. Here an unconstrained optimisation algorithm is used, the Nedler and Mead Symplex method.[8] The MSE is calculated in dollar units, i.e.

$$\$MSE_t = \frac{\sum_{i=1}^{N} (C_{i,t} - \hat{C}_{i,t})^2}{N}, \tag{3.42}$$

where N is the number of options per surface, C is the market's option price (in dollars) and $\hat{C}$ is the estimated price according to different volatility dynamics assumed for the underlying.

The above-mentioned technique yields different option prices by using each time a different volatility assumption. The next step in order to perform forecasting is to average. As a direct consequence of the nonlinear relationship between option price and volatility of the underlying asset however two different averaging approaches are adopted and are expected to yield different results.

The first involves averaging option prices derived from different option models. Each option price is assigned a weight according to its underlying conditional-volatility model which, however varies based on the averaging method considered each time. The second approach involves first averaging individual volatility models and then assign this averaged scheme to the option price. The ultimate objective, however, in both cases is achieving superior forecasting performance amid a spawning literature of single-model pricing approaches.

3.4.3. Methodology for Averaging Option Prices against Averaging the underlying Volatility

In order to achieve consensus when trying to value options while deviating from the assumption of a single model process, there is one more dilemma on which we should focus our attention. This is to say whether it is preferable

[8] Lagarias *et al.* (1998)

to attempt averaging after pricing using different option models or price directly using an already averaged volatility scheme as an assumption.

The first approach can be summarized as follows. Primarily we obtained a collection of different option prices by altering the volatility process of the asset. After the risk-neutral parameters are calibrated and the prices are obtained, the researcher now has under consideration a collection of surfaces (one for every model considered) to average and yield the ultimate prices. The weight assigned to each surface is based on the particular averaging method that is under consideration. Here there are three standard statistical approaches that are taken into (account all explained thoroughly in the volatility averaging section): the Bayesian Model Averaging, the Bayesian Approximation, and different aspects of Thick Modelling Averaging. Each averaging method allocates different weights to each volatility model depending on its assumptions. These are the weights that now will be used to average option prices. It should be pointed out, however, that in simple arithmetic average Thick Modelling we do not actually assign weights, but we take the average of various percentages of the best-performing models. In Thick modelling with Optimization, though, the optimal loadings are determined by the optimisation algorithm and are used to weight volatility models. Similarly, unequal weighting is applied to the volatility models under the Neural Network spectrum.

Regarding the second approach adopted there is no need to estimate different option prices under distinct volatility assumptions. Instead we only estimate prices that assume as the volatility of the underlying security the averaged volatility. This is a constructive as well as pioneer comparison of the two averaging alternatives for option pricing, empirical results of which are discussed in the following sections.

3.5. Data

For the purpose of the analysis the S&P 500 index is selected (source CRSP Database). This is driven by a variety of reasons, mainly due to the fact that it is a highly liquid index and heavily traded. The sample period covers ten years, from 31st August 1997 until 1st September 2007. This yields a total of 2610 observations, which are then divided into the estimation period (in-sample) and the forecasting period (out-of-sample). The first

part involves 9 years or else 2349 observations, while the last 1 year has 261 observations. Afterwards returns are estimated from the index adjusted prices (observations). These data relate to the volatility estimation. The selection of the best single models in-sample in order to use them for the out-of-sample forecasting is made using the Bayesian Information Criterion. Finally the ranking is based on Mean Squared Error

$$MSE_{M_i} = \frac{\sum_{t=1}^{T} \left(h_{t,M_i} - \hat{h}_{t,M_i}\right)^2}{T}, \tag{3.43}$$

where M_i stands for the different single models or averaging methods, h is the actual and $\hat{h}$ is the predicted volatility.

Additionally, as has already been mentioned, volatility is a latent variable and for that reason a proxy is used to estimate its actual value. Usually squared returns are most frequently used to approximate volatility, although this approach is considered as rather noisy and imprecise. For this reason an alternative is adopted, the high-low measure of Bollen and Inder (2002). It is first presented by Parkinson (1980) and later extended by Garman and Klass (1980) and studied thereafter by many researchers. [9] The high-low estimator is defined as follows:

$$\hat{h} = \frac{\left(\ln P_{H_t} - \ln P_{L_t}\right)^2}{4 \ln 2}, \tag{3.44}$$

where P_H is the High Price at time t of S&P 500 and P_L is the Low Price at time t. The choice of this particular proxy is due to the fact that in comparison with high-frequency data, high and low prices are publicly available, easily and freely accessible, while any data manipulation is very quick to perform. Most importantly however the high-low estimator avoids incorporating any biases caused by market microstructure problems (non-synchronous trading, seasonality effects etc.).

Before turning our attention to the option data relevant details, it is important to refer first to some widely observed patterns related to implied volatility. It has long been observed (Rubinstein (1985, 1994), Jackwerth and Rubinstein (1996)) that at-the-money equity options have lower implied volatilities than in- or out-of-the-money options [10]. This phenomenon, also

[9] For more details see Ser-Huang Poon (2005)
[10] Hull (2005)

known as volatility smile or volatility skew, demonstrates the fact that as the strike price increases the volatility decreases (on the contrary, Black and Scholes have predicted that it should stay constant). Volatility smiles help traders to price options. Additionally, an interrelated concept to volatility smile is the volatility term structure. This represents how implied volatility changes for similar options but with different maturities. When we combine the term structure of volatility with the volatility smile, then we generate an implied volatility surface. Volatility surfaces helps us to price options with any strike price and any maturity by indicating the relevant implied volatility level. Figure 3. presents a plot of a random implied volatility surface used in the analysis.

In particular for the option pricing study, a sample of one-year option prices is used instead from 31st August 2006 up to 27th September 2007. The source for the option data is Optionmetrics and it was accessed from Wharton Research Data Services (WRDS). Due to market irregularities, however, the option prices have to be filtered. Following relevant approaches found in literature (Engle *et al.* (2006)), only Wednesday prices are selected, so as to avoid intra-week effects. Also only out-of-the money options are used, due to their liquidity advantage in comparison with in- and at-the-money options. Additionally options with short maturity (less than ten days) or long maturing options (more than 365 days) are discarded. Implied volatility is also an important factor for filtering, such that options with greater than 70% implied volatility are sorted out. It should be pointed out that the price of an option is considered the average of the bid and ask price, and by using this definition a last filter is applied. Options with small value (less than \$0.05) are not included. The above filtering generates 42 distinct surfaces (one for each week) for the estimation period, each containing a variety of different maturing options and with different strike prices. The final total number of filtered option prices is 7384, from which 2508 are call and the rest put options. Table 3.18 summarizes the exact number of options retained in each surface after the filtering process. For brevity the analysis is performed based only on call options. However, it can easily be extended to put options as well.

In order to calibrate the risk-neutral parameters for each particular GARCH-type model that drives the volatility of the underlying the following steps are executed. First, Maximum Likelihood is used to estimate the models'

parameters, and for that reason 2349 available historical returns of S&P 500 (9 years of data) are used as inputs to the optimisation routines. These estimates were afterwards used as starting values to facilitate the calibration procedure of the risk-neutral parameters from the actual option data. Additionally, using Monte Carlo, 20000 simulated option prices are generated for pricing purposes. Finally, regarding the risk-free interest rate required as input to the pricing equation, the US Treasury Bill with maturity three months, six moths and one year is used. In the past various approaches have been adopted in order to accommodate the need to match each individual option maturity to the relevant risk-free rate. For example Engle *et al.* use the method of linear interpolation, while Christofersen *et al.* keeps it constant. Here we have observed a small divergence between the three maturities available. For this reason a simple average of the three is used as an input when pricing. In other words, at each step the average of the month's TB rates for all three maturities is estimated for use in the pricing equations.

3.6. Empirical results

3.6.1. Empirical results for Volatility Forecasting

The idea of testing the superiority of a single GARCH-type model against competing averaged schemes involves initially determining the optimal lags related to each single volatility model. For that reason the BIC model selection criterion is used in order to access the in-sample performance of the competing alternatives (see Table 3.3.). According to this measure, the same optimal single model selection should outperform during the estimation as well as during the forecasting period.

In more detail, when the S&P 500 index is used the BIC criterion indicates that the overwhelming majority of the best single models has specifications of order one (i.e. (p,q)=(1,1)). On the contrary, according to the same model selection indicator only Fattailed GARCH differs and is specified to have (2,1) lags instead. Afterwards we are left with ten competing single models which will be used onwards for comparing and contrasting purposes against averaging methods. As mentioned in the previous section, the error is calculated based on high-low estimator as a proxy of the actual volatility,

while the error measure is the Mean Squared Forecasting Error (MSFE). Different error measures could also be selected, but the MSFE is a common choice in relevant research papers.

Together with the MSFE an additional performance indicator, econometric this time, is used, the Diebol-Mariano test (Tables 3.4., 3.5.). In addition Table 3.2. shows the final ranking of all 27 single and averaged models used to produce an optimal out-of-sample forecast. According to that table, we notice the following. All the first ranking positions (up to the 11th place) are occupied by averaging proposals (with only exception the TGARCH model). Noteworthy is also the fact that on average single conditional-volatility models are poor performers, demonstrating the highest forecasting error. The importance of this results stems from the fact that the performance of the competing models is reported for a lengthy period of time (the-out-of sample period covers a year). In other words, empirical results indicate that the investor would have gained better predictive supremacy by using an averaged model. The driving force behind this outcome is model uncertainty. This can not be avoided in advance if only a single model is selected using an in-sample error measure, and thus superior predictability is in jeopardy. This is because usually time horizon, data frequencies and financial assets classes differ according to the researcher's needs. This in return implies that it is practically impossible to determine empirically which is the one and only alternative among a vast number of model choices which can always outperform.

There are of course possible explanations of the shortcomings related to single conditional-volatility models and their inability to accurately predict the future. One of these is the way the best model is selected, since someone has to entirely rely on the in-sample performance of all competing choices with all different lag specifications available to complicate matters further. This can easily lead to a suboptimal decision, since out-of-sample requirements can change, the environment could be much more treacherous and consequently what has been selected as a best choice in the estimation period might be rendered suboptimal out-of-sample.

Despite accounting for model uncertainty there is an additional advantage of averaging. That is the fact that even the performance of the best single model out-of-sample (identified accurately or not in advance by the researcher using any model selection criterion) can be enhanced when at-

tempting averaging. This is exactly the reason why we find an averaging proposal still ranked in the first place when all estimated models are compared. In the appendix Table A.2 provides relevant information of the out-of-sample MSE. Here the total 187 single and averaged volatility models (and not just the best in-sample using BIC) are ranked from the lowest to the highest according to MSE. In the first position we find Thick Modelling with Optimization and 20% retention percentage, yielding the lowest on average out-of-sample error. As when we compared 27 volatility forecasting alternatives, all averaging methods are ranked at the first positions. Last, around fifty positions are occupied by single models, which again substantiates the claim that their majority has an inferior forecasting performance. There is however an exception among them, the Threshold GARCH, but as is clear from the analysis the researcher could not have foreseen this in advance. In other words, no one could guaranteed that the Threshold GARCH in any given circumstances can still produce better out-of-sample results and that the lags chosen by optimizing in-sample an information criterion or another error measure are the optimal lags for some other empirical experiment.

Moving to table 3.4 and 3.5, we find results for the Diebold-Mariano test. This test is conducted in order to substantiate econometrically the superior performance of one method against another. This is a popular measure used to examine whether the average difference of the loss function of two competing models is non-zero. Which one outperformed is determined by the value of a t-statistic. For 95% confidence interval, if DM t-statistic is $>$ 1.96 then the losses of the first model are significantly larger than those of the second one and if DM t-statistic $<$ -1.96 the second model has higher losses than the first one. The loss function used for the needed difference to be tested is once more the MSE. Although at a first look dominance of a particular model-averaging method is not apparent, closer examination reveals that some averaging methods do seem to outperform more frequently than other competing single proposals.

Starting with the analysis of the results for the single volatility models, the worst of them is Fattailed GARCH, which is inferior to every other model and only twice equal to its competing alternatives. Regarding the other single models, 7 out of 10 outperform three times volatility forecasting shemes and only 1 out of 10 models (the T-GARCH in particular) offers

four times better forecasting accuracy. Moving to averaging methods, a more evident outperformance can be observed. Here 6 averaging techniques dominate their alternatives four times (in contrast with single models of which only one has such a performance), while 3 models have five times better predictive power. All the above discussed can be regarded as a clear argument in favor of averaging, now using econometric criteria as well.

3.6.2. Empirical results for Option Pricing

Some additional comments regarding averaging under the option pricing spectrum have to be provided here. These relate to the weights that will be assigned to each model alternative and their origin in order to perform averaging.

When we average option prices based on models with different assumptions about the volatility dynamics of the underlying, two approaches are introduced related to the weights assigned to these distinct models. Initially we have available the weights already calculated when performing averaging of the conditional-volatility from the previous section. For brevity of analysis here the loadings available only for the methods of Bayesian Model Averaging, Bayesian Approximation and Thick modelling with simple average and with retention percentages from 10% to 50% will be used to average option prices. In addition an alternative weighting method is added in order to determine those loadings that each model will receive. This method, however, relates only to the Thick Modelling and the Bayesian Approximation approach.

Until now BIC has been used solely for model selection and averaging purposes. However, this dependence could potentially be suboptimal when performing option pricing. Instead we could opt to minimize another error measure, is the in-sample $RMSE of the option prices. The importance of this approach is that now the focal point becomes the pricing ability of each model based on an error measure, and not on the BIC from the volatility analysis.

Regarding the method used to select the single option models that are compared against averaging schemes, we rely again on the BIC. In other words, the single option models that are used are those that assume as the underlying volatility dynamics the process that is selected as superior in

the volatility analysis section (using the BIC criterion as model selection indicator). Finally the results can be summarized as follows.

According to standard practice,[11] one week ahead is selected as an out-of-sample forecasting target for a whole one-year horizon. This in turn suggests that after a surface is calibrated and the risk-neutral parameters are acquired from optimizing the fitting of the pricing model to the actual option prices, these parameters are used to forecast next week's surface prices using Monte Carlo simulation. In table 3.7 the relevant results can be observed, where as mentioned before single option pricing models have specifications determined in the first part of the analysis by minimizing the in-sample BIC criterion of the underling's volatility process. These are the models that an objective researcher would use if he should determine in the estimation period which are the dominant ones according to an error measure (BIC in this case).

As already observed in the volatility sectionS the investor should decide for one and only one optimal process to use for forecasting option prices. However, the analysis is indeed extended to all available models, those that are selected in-sample and those not, so as to substantiate the catholic averaging superiority. Table A.3 in the appendix summarizes the performance of all models considered in this research collectively, including those single proposals that were sorted out in the estimation period as suboptimal. Results regarding this collective table demonstrate the following. In all first five places we find Thick Modelling with weights determined based on optimizing the in-sample option pricing error and under the "Averaged Option Prices" scheme (and not with the "Averaged Volatility" alternative). In general TM is preferable for smaller "retention" percentages of the best performing models. At sixth place, however, we find a single alternative, GJR-GARCH(3,1), not one of those that has been picked during the estimation period based on BIC criterion.

According to Table 3.7. now which is the main focus of the results, rankings reveal that the vast majority of the averaged option pricing methods have better forecasting power than their single counterparts. Thick Modelling with weights calculated optimizing the option $RMSE under the "Averaged Volatility" exhibits satisfactory overall performance, although

[11] See papers of Barone-Adesi, Egle, Mancini (2006), Christoffersen, Jacobs (2004), Heston and Nandi (2000).

the most dominant of all as mentioned in the cumulative table as well, is Thick Modelling under the "Averaging Models" scheme. In general the method works better when weighting is based on the option $RMSE and not on the loadings determined in the volatility section using the BIC measure.

The overall superior performance of the averaging suggestions is fraught with significance in this section for one additional reason. The target date forecasts relate to is a week later than where we stand today. The error reported offers to the potential investor the choice to avoid the recalibration of parameters during this period, which is important for larger portfolios. In other words, the averaging methods reassure the realization of a positive outcome of an investment even a week later after first calibrating the parameters for the pricing equation.

A final point that has also been raised shortly before is that all averaging methods do not seem to perform equally well. This is natural to expect as for some approaches the reduced forecasting potential has a underlying interpretation. Those weights generated from the conditional-volatility forecasting section and used here as well for averaging option pricing relate to models whose parameters were derived under the subjective measure (real-world data were used). This could be a reason to affect the overall performance of those averaging suggestions that are based on volatility weights alone.

Other averaging methods, however, such as the Bayes Approximation (BA) under the "Averaged Volatility" approach with weighting based both on $RMSE and on BIC, exhibited a similar performance, which is above 94% of all models in general in the cumulative table A.3 and in the top ten performers in Table 3.7 Bayesian Model Averaging (BMA) also performs better when we have "Averaged Volatility" method applied and not the "Averaged Models" alternative. This performance can be attributed to the fact that these two methods (BA and BMA) consider weighting of all single models to which relevant posterior probabilities are assigned. Thick Modelling, on the other hand, accounts only for a specific percentage of the total, which includes models better performing in the estimation period. It is natural consequently to deduce that there might be models that perform adequately well out-of-sample and others not, which were however optimal in-sample but not afterwards, and drive the average performance of BMA and BA downwards.

3.6.2.1. Empirical results when Comparing with a Benchmark

The above pricing methods are also compared to Duan's model, which uses an analytical formula to price options (mentioned at the beginning of the option analysis). The benchmark model seriously underperforms all averaged methods for the out-of-sample period as well as single models, as revealed by the average $RMSE for all 42 surfaces for which relevant forecasts are made.

3.6.2.2. Empirical results according to Maturity and Moneyness

Tables 3.8 until 3.15 include out-of-sample results when forecasting option prices, divided however into different moneyness and maturity categories. Additionally, cumulative tables of all models considered in this analysis, although available, are not reported in the thesis for brevity. Analysis is restricted to the single models selected using the BIC as an in-sample performance measure. The reason for performing this subcategorization of results (according to the maturity and moneyness) is due to the fact that that some averaging methods may work better under different horizons and maturities, so we have to identify these tendencies wherever they exist.

Beginning with Table 3.8, only call options with maturity less than 60 days and with moneyness less than 1.05 are included. Here there is a perceptible predictive supremacy of Thick Model Averaging under the "Averaged Volatility" spectrum and with weighting based on the option $RMSE. Additionally it is advisable to select retention smaller percentages of the best models in general when applying the method, although the performance between them does not vary significantly. However, Thick Modelling with weights, specified using the BIC both for the "Averaged Volatility" and "Averaged Models" approach, is a suboptimal forecasting choice. Regarding BMA and BA, "Averaged Volatility" is still preferable, which is in general the main conclusion to be drawn from this subcategory. To summarize, TM for short-term maturity options with small moneyness has the best forecasting power, while merging volatilities and not option prices should be chosen for all averaging methods.

Moving to Table 3.9, which refers only to options with maturity less than 60 days and moneyness between 1.05 and 1.15, we have almost identical results. Options with "Averaged Volatility" and Thick Modelling prevail,

ranked for the various retention percentages at the first five places. Additionally BMA and BA generate more accurate forecasts under the "Averaged Volatility" perspective.

When the maturity of the options increases, however, the overall results do not remain the same. Table 3.10. incorporates those rankings related to options with maturities from 60 to 160 days and moneyness less than 1.05. Now the TM with "Averaged Models" and weighting according both to option $RMSE and BIC is preferable to the "Averaged Volatility" equivalent methods. The same holds for BMA and BA: "Averaged Models" processes outperform their rival alternative, while on the whole the two methods strengthen their predictive power. To summarize, for larger maturities the method of option model averaging instead of an averaged volatility dynamic provides lower $RMSE on average for the entire out-of-sample period for all different approaches. Conspicuous is also the fact that BMA demonstrates a stronger forecastability.

To highlight how the positioning changes when moneyness is increased (maturity horizon remains between 60 to 160 days), we need the next Table, 3.11. TM with weighting based on option $RMSE under the "Averaged Models" method is once more at the forefront as the best predicting model. Again BMA and BA give lower error under the "Averaged Volatility" perspective. All in all "Averaged Volatility" method for this surface demonstrates good performance for the various averaging schemes. For the deep out-of-the money options (now moneyness is larger than 1.15), table 3.12, provides relevant results. TM with BIC based weighting still under the "Averaged Models" method and for all retention percentages is ranked at the top.

Table 3.13 has options maturing in the long run but with smaller moneyness (<1.05). TM both with "Averaged Models" and "Averaged Volatility" but with weights determined by optimizing the in-sample option $RMSE exhibit superior predictability. On the other hand both BMA and BA, using BIC and option error for weighting under the "Averaged Volatility" spectrum, have better forecasting power. Similar results are reported for options with the same long term maturity but for moneyness from 1.05 to 1.15 (Table 3.14). The TM with with "Averaged Models" and $RMSE is superior over the other averaging methods. Finally, for even more increased moneyness (>1.15, Table 3.125), TM both with "Averaged Models" and

"Averaged Volatility" under $RMSE weighting scope is still the ultimate dominant performer.

For the above extensive analysis some general conclusions can be drawn. Thick Modelling under the option forecasting error criterion for weighting purpose demonstrates the most powerful predictive ability, both under the "Averaged Models" as well as under the "Averaged Volatility" formation. Especially "Averaged Models" for TM works better for deep out-of-the-money options with long maturity and for close at-the-money options but with shorter maturity. On the other hand, TM but with BIC-based weights has inferior predictive dynamics for most maturities and moneyness levels, with only the exception of out-of-the-money options with average maturity horizon.

3.7. Conclusions

This chapter is an attempt to address volatility model uncertainty and its direct implications for option pricing using averaging techniques. In particular, averaging is performed under three distinct approaches: Bayesian Model Averaging first, Bayesian Approximation second and finally Thick Model Averaging. Along with the novelty of utilizing Bayes theory to perform averaging based on the models' posterior probabilities, two new approaches of Thick Modelling are introduced. The first is Thick Modelling with Optimization for the weighting and finally a simple Feed Forward Neural Network to attempt the same task of tracking optimal loadings necessary for averaging. This exhibits forecasting supremacy among other 187 competing alternatives. On the other hand TM using Neural Networks did not met expectations. This can be attributed to the fact that the initial assumption of non-linear relationships that drive the dynamics between different models could not be validated empirically.

The second part utilizes the result of the volatility forecasts to price options. Here averaging is applied to a considerable number of single option pricing models which were constructed by changing the assumption of the volatility process of the underlying. However, the observed non-linear relationship between volatility and option prices led us to take into account a twofold approach for averaging. The first one involves "Averaged Models" using the various option pricing equations, and the second approach

relates to the "Averaged Volatility" where just one single volatility process, the averaged volatility, is fed into the pricing formula. Additionally a new suggestion for determining the weights is used; this which involves the optimisation of the in-sample $RMSE of option prices to identify the loadings.

Results indicate first that TM with BIC weighting does not meet performance expectations. This is because weights were initially determined under the objective measure using real-world observations and as a result directly applying those weights to option models proved to be suboptimal. However, BMA retaines its predictive power as an averaging alternative offering satisfactory low forecasting error using both the BIC and the option $RMSE. Overall, TM with weights specified using option $RMSE demonstrated the strongest results.

Regarding which of the two averaging suggestions empirical results indicate as the most dominant to forecast option prices we have to notice the following. The "Averaged Models" proposal for pricing is on average more competent than the "Averaged Volatility" method. Especially for deep out-of-the-money options with maturities larger than 160 days and out or near at the money options for smaller maturity horizons, we observed that "Averaged Models", together with TM and weighting based on option error, performed better than any other alternative.

Key Code	Model	Acronym
1	GARCH	-
2	APGARCH	-
3	AVGARCH	-
4	NGARCH	-
5	NAGARCH	-
6	TGARCH	-
7	GJRGARCH	-
8	FATGARCH	-
9	EGARCH	-
10	FIGARCH	-
11	Bayes Approximation	BA
12	Bayesian Model Averaging	BMA
13	TM simple average with 10% best models	TM 10%
14	TM simple average with 20% best models	TM 20%
15	TM simple average with 30% best models	TM 30%
16	TM simple average with 40% best models	TM 40%
17	TM simple average with 50% best models	TM 50%
18	TM with optimization with 10% best models	TM 10% OPT
19	TM with optimization with 20% best models	TM 20% OPT
20	TM with optimization with 30% best models	TM 30% OPT
21	TM with optimization with 40% best models	TM 40% OPT
22	TM with optimization with 50% best models	TM 50% OPT
23	TM with Neural Networks with 10% best models	TM 10% NN
24	TM with Neural Networks with 20% best models	TM 20% NN
25	TM with Neural Networks with 30% best models	TM 30% NN
26	TM with Neural Networks with 40% best models	TM 40% NN
27	TM with Neural Networks with 50% best models	TM 50% NN

Table 3.1.: Model Key Indicators and acronyms for Volatility tables.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0	3.32435E-09	TM 20% OPT	**15**	1	1	3.62612E-09	NAGARCH
2	1	1	3.3442E-09	TGARCH	**16**	1	1	3.68865E-09	FIGARCH
3	0	0	3.35561E-09	TM 10% OPT	**17**	0	0	3.68909E-09	BMA
4	0	0	3.36393E-09	TM 20%	**18**	1	1	3.78633E-09	NGARCH
5	0	0	3.43662E-09	TM 30% OPT	**19**	1	1	3.81868E-09	EGARCH
6	0	0	3.44233E-09	TM 30%	**20**	1	1	3.87758E-09	GARCH
7	0	0	3.46124E-09	TM 50%	**21**	2	1	3.9277E-09	FATGARCH
8	0	0	3.46125E-09	TM 50% OPT	**22**	1	1	3.97704E-09	GJRGARCH
9	0	0	3.47887E-09	TM 40% OPT	**23**	0	0	4.32081E-09	TM 50% NN
10	0	0	3.47983E-09	TM 40%	**24**	0	0	4.57424E-09	TM 40% NN
11	0	0	3.48288E-09	TM 10%	**25**	0	0	6.18459E-09	TM 20% NN
12	1	1	3.50788E-09	APGARCH	**26**	0	0	6.27352E-09	TM 30% NN
13	0	0	3.60817E-09	BA	**27**	0	0	7.62863E-09	TM 10% NN
14	1	1	3.60995E-09	NGARCH					

Table 3.2.: Rankings according to RMSE of averaging methods and only those volatility models selected according to the best in-sample BIC criterion.

Model							Best Model (BIC)
EGARCH		**Lag p**					
	Lag q	**0**	**1**	**2**	**3**	**4**	
	0	0	-16070.1	-16175.1	-5672.69	-2806.77	
	1	-16088	-16705.3	5490.348	1052.017	-2692.82	**[1,1]**
	2	-16142.2	0	-10522.5	2000055	4866.206	
	3	-16146.9	-16139.1	4857.885	-2876.41	2000071	
	4	723.7795	4855.113	4862.981	2000071	4874.342	
AVGARCH		**Lag p**					
	Lag q	**0**	**1**	**2**	**3**	**4**	
	0	0	0	0	0	0	
	1	161028.3	-16679	-16671.1	-16663.3	-16659.7	**[1,1]**
	2	147116.5	-16671.1	-16665.8	-16663.4	-16660.1	
	3	146889.2	-16663.3	-16657.9	-16657.3	-16651.9	
	4	146698.6	-16655.4	-16649.5	-16649.4	-16644.4	
GARCH		**Lag p**					
	Lag q	**0**	**1**	**2**	**3**	**4**	
	0	0	-16009.3	-16001.4	-15993.5	-15985.7	
	1	-16126.9	-16613.3	-16547.5	-16544.9	-16527.4	**[1,1]**
	2	-16250.1	NaN	-16578.7	-16574.1	-16580.7	
	3	-16298.5	-16590.3	NaN	-16569.3	-16577.6	
	4	-16371.3	-16501.1	NaN	-16538.2	-16508.5	
NGARCH		**Lag p**					
	Lag q	**0**	**1**	**2**	**3**	**4**	
	0	0	0	0	0	0	
	1	4831.367	-16605.5	-16597.6	-16588.7	-16581.7	**[1,1]**
	2	4839.234	-16600.6	-16598.6	-16589.2	-16586.3	
	3	4847.101	-16592.8	-16590.5	-16584.6	-16578.9	
	4	4854.968	-16584.9	-16582.8	-16576.7	-16571.3	
GJRGARCH		**Lag p**					
	Lag q	**0**	**1**	**2**	**3**	**4**	
	0	0	-16734.4	-16726.8	-16724.5	-16722.4	**[0,1]**
	1	-16147.6	-16726.6	-16718.9	-16716.7	-16714.5	
	2	-16262.8	-16718.7	-16711	-16708.8	-16706.6	
	3	-16318.4	-16710.8	-16703.2	-16700.9	-16698.8	
	4	-16385.8	-16703	-16695.3	-16693.1	-16690.9	

APGARCH	**Lag p**					
	Lag q	**0**	**1**	**2**	**3**	**4**	
	0	0	0	0	0	0	
	1	4836.655	-16735.6	-16728.6	-16727.3	-16725.9	**[1,1]**
	2	4844.522	-16723	-16720.7	-16719.4	-16717.6	
	3	4852.389	-16718.1	-16712.8	-16711.2	-16703.6	
	4	4860.256	-16712.2	-16705.1	-16703.7	-16702.2	
FATGARCH	**Lag p**					
	Lag q	**0**	**1**	**2**	**3**	**4**	
	0	0	-16009.3	-16001.4	-15993.4	-15985.7	
	1	-16126.9	-16555.3	-16547.4	-16539.7	-16532.8	
	2	-16250.1	-16578.6	-16577.9	-16574.8	-16572.1	**[2,1]**
	3	-16298.5	-16570.7	-16570	-16567	-16564.8	
	4	-16371.3	-16563.6	-16562.3	-16561.8	-16561.9	
NAGARCH	**Lag p**					
	Lag q	**0**	**1**	**2**	**3**	**4**	
	0	0	0	0	0	0	
	1	82924.39	-16709.7	-16701.8	-16697.3	-16693	**[1,1]**
	2	75972.42	-16701.8	-16697.4	-16699.4	-16696.3	
	3	75862.68	-16693.9	-16689.5	-16691.6	-16688.4	
	4	75771.33	-16686.1	-16681.7	-16683.7	-16680.6	
TGARCH	**Lag p**					
	Lag q	**0**	**1**	**2**	**3**	**4**	
	0	0	109942.3	74450.11	62244.32	54249.82	
	1	157332.8	-16743.6	-16736.7	-16735.7	-16734.4	**[1,1]**
	2	147034	-16735.7	-16728.8	-16670.1	-16726.6	
	3	146873.3	-16727.9	-16720	-16662.2	-16712.1	
	4	146694.4	-16720.5	-16713.4	-16712.1	-16704.3	

Table 3.3.: Table of the in-sample best performing models according to BIC criterion.All the GARCH-type models are examined up to lag 4. The best performing according to BIC information criterion is the one with the lowest BIC value. The last column gives in the squared brackets the (p,q) lag of this best model.

11	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
2	0.03	0.44	0.59	0.23	0.07	0.55	0.00	0.81	0.33	0.25	0.36	0.01	0.01	0.02	0.04	0.07	0.08	0.01	0.02	0.04	0.07	0.00	0.22	0.00	0.27	0.25
3		0.64	0.00	0.15	0.34	0.06	0.00	0.20	0.41	0.28	0.13	0.73	0.09	0.25	0.60	0.56	0.42	0.06	0.22	0.59	0.56	0.00	0.17	0.00	0.08	0.07
4			0.45	0.93	0.01	0.40	0.00	0.44	0.79	0.99	0.69	0.53	0.27	0.40	0.49	0.33	0.01	0.19	0.38	0.49	0.33	0.00	0.21	0.00	0.18	0.17
5				0.22	0.03	0.39	0.00	0.87	0.62	0.18	0.47	0.00	0.00	0.00	0.01	0.02	0.02	0.00	0.00	0.01	0.02	0.00	0.21	0.00	0.20	0.19
6					0.03	0.23	0.00	0.43	0.80	0.76	0.48	0.09	0.01	0.02	0.04	0.01	0.12	0.01	0.02	0.04	0.01	0.00	0.20	0.00	0.14	0.15
7						0.10	0.00	0.05	0.23	0.01	0.01	0.36	0.92	0.56	0.37	0.29	0.94	0.92	0.58	0.38	0.29	0.00	0.17	0.00	0.09	0.06
8								0.56	0.26	0.25	0.30	0.04	0.04	0.06	0.08	0.10	0.10	0.04	0.06	0.08	0.10	0.00	0.24	0.00	0.37	0.36
9									0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.04	0.00	0.00	0.00
10									0.45	0.41	0.50	0.08	0.13	0.15	0.18	0.16	0.03	0.11	0.15	0.18	0.16	0.01	0.23	0.00	0.33	0.17
11										0.76	1.00	0.25	0.19	0.28	0.36	0.37	0.11	0.16	0.27	0.36	0.37	0.00	0.17	0.00	0.18	0.02
12											0.41	0.22	0.02	0.05	0.08	0.00	0.11	0.01	0.04	0.07	0.00	0.00	0.20	0.00	0.14	0.15
13												0.02	0.05	0.07	0.08	0.03	0.06	0.04	0.06	0.08	0.03	0.00	0.22	0.00	0.21	0.18
14													0.33	0.65	0.97	0.82	0.43	0.24	0.61	0.96	0.82	0.00	0.17	0.00	0.10	0.05
15														0.08	0.07	0.31	0.97	0.04	0.11	0.08	0.31	0.00	0.15	0.00	0.03	0.04
16															0.08	0.76	0.62	0.04	0.05	0.08	0.76	0.00	0.16	0.00	0.06	0.05
17																0.70	0.46	0.04	0.05	0.09	0.70	0.00	0.17	0.00	0.08	0.07
18																	0.48	0.17	0.69	0.71	0.50	0.00	0.17	0.00	0.09	0.08
19																		0.88	0.64	0.46	0.48	0.00	0.15	0.00	0.07	0.03
20																			0.04	0.04	0.17	0.00	0.14	0.00	0.03	0.04
21																				0.05	0.69	0.00	0.16	0.00	0.06	0.05
22																					0.71	0.00	0.17	0.00	0.08	0.07

23	0.00	0.17	0.00	0.09	0.08
24	0.30	0.29	0.00	0.00	0.01
25	0.96	0.31	0.26		
26	0.00	0.04	0.00		
27	0.72				
	0.00				

Table 3.4.: Table of the p-values generated by the Diebold-Mariano test.

The Diebold-Mariano test is executed using the best performing in-sample volatility models according to the BIC criterion and all available averaging approaches. Each model is compared to the remaining volatility models as well as with all the averaging methods. In the test the mean of the difference between MSE out-of-sample is used to determine whether one model is superior to another. Numbers 1 to 27 both in the first vertical and horizontal line indicate the model codes. The numbers in the shells are the p-values of the t-statistic at 95% confidence interval. Key indicators for models can be found in Table 3.1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1	-	-
2		-	2	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	2	-	-
3			-	-	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	3	-	-
4				-	-	-	-	4	-	-	-	-	13	14	15	16	-	-	19	20	21	-	4	-	4	-	-
5					-	-	-	5	-	-	-	-	-	-	-	-	17	-	19	-	-	22	5	-	5	-	-
6						-	-	6	-	-	-	6	-	-	-	-	-	-	-	-	-	-	6	-	6	-	-
7							-	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7	-	7	-	-
8									9	10	11	12	13	14	15	16	17	18	19	20	21	22	-	-	25	26	27
9										-	-	-	-	-	-	-	-	-	-	-	-	-	9	-	9	-	-
10											-	-	-	-	-	-	-	-	-	-	-	-	10	-	10	-	-
11												-	-	-	-	-	17	-	19	-	-	22	11	-	11	-	-
12													-	-	-	-	-	-	-	-	-	-	12	-	12	-	-
13														-	-	-	-	-	-	-	-	-	13	-	13	-	-
14															-	-	-	-	-	-	-	-	14	-	14	-	-
15																-	-	-	-	-	-	-	15	-	15	-	-
16																	-	-	-	-	-	-	16	-	16	-	-
17																		-	-	-	-	-	17	-	17	-	-
18																			-	-	-	-	18	-	18	-	-
19																				-	-	-	19	-	19	-	-
20																					-	-	20	-	20	-	-
21																						-	21	-	21	-	-

22																														22	-	22	-	-
23																															-	-	26	27
24																																-	-	-
25																																	-	27
26																																		-
27																																		

Table 3.5.: Table including Diebold-Mariano comparison results.

The table shows which model has better predictive power than its counterpart. " - " indicates that the pair of competing models or averaging methods has statistically equal forecasting dynamics, while the encounter of a number in a row implies the model that dominated when compared against another alternative.

Key Code	**Model**	Acronym
1	GARCH	-
2	APGARCH	-
3	AVGARCH	-
4	NGARCH	-
5	NAGARCH	-
6	TGARCH	-
7	GJRGARCH	-
8	FATGARCH	-
9	EGARCH	-
10	FIGARCH	-
	Options Model Averaging	
11	Bayes Approximation	BA
12	Bayesian Model Averaging	BMA
13	TM simple average with 10% best models	TM 10%
14	TM simple average with 20% best models	TM 20%
15	TM simple average with 30% best models	TM 30%
16	TM simple average with 40% best models	TM 40%
17	TM simple average with 50% best models	TM 50%
18	Bayes Approximation with option error	BA OPT
19	TM simple average with option error with 10% best models	TM 10% OPT
20	TM simple average with option error with 20% best models	TM 20% OPT
21	TM simple average with option error with 30% best models	TM 30% OPT
22	TM simple average with option error with 40% best models	TM 40% OPT
23	TM simple average with option error with 50% best models	TM 50% OPT
	Options with averaged Volatility	
24	Bayes Approximation	BA VOL
25	Bayesian Model Averaging	BMA VOL
26	TM simple average with 10% best models	TM 10% VOL
27	TM simple average with 20% best models	TM 20% VOL
28	TM simple average with 30% best models	TM 30% VOL
29	TM simple average with 40% best models	TM 40% VOL
30	TM simple average with 50% best models	TM 50% VOL
31	Bayes Approximation with option error	BA OPT VOL
32	TM simple average with option error with 10% best models	TM 10% OPT VOL
33	TM simple average with option error with 20% best models	TM 20% OPT VOL

34	TM simple average with option error with 30% best models	TM 30% OPT VOL
35	TM simple average with option error with 40% best models	TM 40% OPT VOL
36	TM simple average with option error with 50% best models	TM 50% OPT VOL

Table 3.6.: Model Key Indicators and acronyms for Option tables.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0	1.8347	TM 10% OPT	19	0	0	1.9448	BMA VOL
2	0	0	1.8450	TM 20% OPT	20	0	0	1.9864	TM 50% VOL
3	0	0	1.8533	TM 30% OPT	21	0	0	1.9939	TM 10%
4	0	0	1.8619	TM 40% OPT	22	0	0	2.0084	TM 40% VOL
5	0	0	1.8708	TM 50% OPT	23	1	1	2.0246	GARCH
6	0	0	1.9097	TM 30% OPT VOL	24	0	0	2.0429	TM 30% VOL
7	1	1	1.9108	AVGARCH	25	0	0	2.0538	TM 20% VOL
8	0	0	1.9136	TM 40%	26	0	0	2.0773	TM 10% VOL
9	0	0	1.9149	BA VOL	27	1	1	2.1456	NGARCH
10	0	0	1.9150	BA OPT VOL					
11	0	0	1.9190	TM 50%	28	2	1	4.2148	FATGARCH
12	0	0	1.9193	TM 20% OPT VOL	29	0	0	4.2963	BMA
13	0	0	1.9199	TM 40% OPT VOL	30	0	0	5.3178	BA OPT
14	0	0	1.9217	TM 50% OPT VOL	31	1	1	7.5152	GJRGARCH
15	1	1	1.9276	NAGARCH	32	1	1	9.6901	TGARCH
16	0	0	1.9285	TM 10% OPT VOL	33	0	0	11.8704	BA
17	0	0	1.9310	TM 30%	34	1	1	15.4674	APGARCH
18	0	0	1.9334	TM 20%					

Table 3.7.: Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0	1.0293	TM 10% OPT VOL	**17**	1	1	1.5869	GARCH
2	0	0	1.0712	TM 20% OPT VOL	**18**	0	0	1.6281	TM 10% VOL
3	0	0	1.0733	TM 40% OPT VOL	**19**	1	1	1.6739	NGARCH
4	0	0	1.0863	TM 30% OPT VOL	**20**	0	0	1.7576	TM 40%
5	0	0	1.0991	TM 50% OPT VOL	**21**	0	0	1.8356	BMA
6	0	0	1.2970	BMA VOL	**22**	0	0	1.8666	BA OPT
7	0	0	1.3287	BA OPT VOL	**23**	0	0	1.8678	BA
8	0	0	1.3294	BA VOL	**24**	0	0	1.8845	TM 20%
9	0	0	1.3972	TM 10% OPT	**25**	0	0	1.9559	TM 30%
10	1	1	1.4153	AVGARCH	**26**	0	0	2.1484	TM 10%
11	1	1	1.4556	NAGARCH	**27**	1	1	2.6054	TGARCH
12	0	0	1.4660	TM 20% OPT	**28**	0	0	2.8636	TM 30% VOL
13	0	0	1.4687	TM 20% VOL	**29**	0	0	3.2183	TM 40% VOL
14	0	0	1.4957	TM 30% OPT	**30**	0	0	4.3254	TM 50%
15	0	0	1.5271	TM 40% OPT	**31**	0	0	5.3164	TM 50% VOL
16	0	0	1.5840	TM 50% OPT	**32**	1	1	19.3102	GJRGARCH

Table 3.8.: Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity less than 60 and Moneyness less than 1.05.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0	0.2552	TM 10% OPT VOL	**17**	1	1	0.5729	GARCH
2	0	0	0.2912	TM 20% OPT VOL	**18**	0	0	0.6030	TM 40%
3	0	0	0.2992	TM 30% OPT VOL	**19**	0	0	0.6080	TM 20% VOL
4	0	0	0.3102	TM 40% OPT VOL	**20**	0	0	0.6232	BMA
5	0	0	0.3416	TM 50% OPT VOL	**21**	0	0	0.6919	TM 10% VOL
6	1	1	0.4111	AVGARCH	**22**	1	1	0.6991	NGARCH
7	0	0	0.4124	TM 10% OPT	**23**	0	0	0.7148	BA OPT
8	1	1	0.4156	NAGARCH	**24**	0	0	0.7152	BA
9	0	0	0.4228	BMA VOL	**25**	0	0	0.8408	TM 20%
10	0	0	0.4646	TM 20% OPT	**26**	0	0	0.8641	TM 30% VOL
11	0	0	0.4737	BA OPT VOL	**27**	0	0	0.8710	TM 40% VOL
12	0	0	0.4740	BA VOL	**28**	0	0	0.8729	TM 30%
13	0	0	0.4848	TM 30% OPT	**29**	0	0	0.9157	TM 10%
14	0	0	0.5126	TM 40% OPT	**30**	0	0	1.1608	TM 50% VOL
15	1	1	0.5334	TGARCH	**31**	0	0	2.2471	TM 50%
16	0	0	0.5480	TM 50% OPT	**32**	1	1	11.1623	GJRGARCH

Table 3.9.: Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity less than 60 and Moneyness from 1.05 to 1.15.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0	1.7007	TM 20%	**17**	0	0	2.1441	BMA VOL
2	1	1	1.7831	AVGARCH	**18**	0	0	2.1562	BA OPT
3	0	0	1.7856	TM 50% OPT	**19**	0	0	2.1567	TM 30%
4	0	0	1.7865	BMA	**20**	0	0	2.1569	BA OPT VOL
5	0	0	1.8052	BA	**21**	0	0	2.3049	TM 50% OPT VOL
6	0	0	1.8053	BA VOL	**22**	0	0	2.3704	TM 40% OPT VOL
7	0	0	1.8083	TM 40% OPT	**23**	0	0	2.3796	TM 20% OPT VOL
8	1	1	1.8380	NGARCH	**24**	0	0	2.4015	TM 30% OPT VOL
9	1	1	1.8529	NAGARCH	**25**	0	0	2.4622	TM 10% OPT VOL
10	0	0	1.8589	TM 30% OPT	**26**	0	0	4.3596	TM 40%
11	1	1	1.8654	GARCH	**27**	0	0	5.0389	TM 40% VOL
12	0	0	1.8818	TM 10% OPT	**28**	0	0	5.0920	TM 30% VOL
13	0	0	1.8822	TM 20% OPT	**29**	1	1	5.9657	TGARCH
14	0	0	1.9466	TM 10% VOL	**30**	0	0	6.6062	TM 50%
15	0	0	1.9480	TM 10%	**31**	1	1	9.2409	GJRGARCH
16	0	0	1.9543	TM 20% VOL	**32**	0	0	9.4469	TM 50% VOL

Table 3.10.: Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity from 60 to 160 and Moneyness less than 1.05.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0	0.7089	TM 40% OPT	**17**	0	0	0.9147	TM 50% OPT VOL
2	1	1	0.7094	AVGARCH	**18**	0	0	0.9797	TM 10% VOL
3	0	0	0.7220	TM 30% OPT VOL	**19**	1	1	1.0051	GARCH
4	0	0	0.7251	TM 20% OPT VOL	**20**	0	0	1.0241	BMA
5	0	0	0.7320	TM 50% OPT VOL	**21**	0	0	1.0259	BA OPT
6	0	0	0.7587	TM 10% OPT VOL	**22**	0	0	1.0262	BA
7	0	0	0.8197	BA OPT VOL	**23**	0	0	1.0895	TM 20%
8	0	0	0.8199	BA VOL	**24**	0	0	1.1469	TM 30%
9	0	0	0.8206	TM 10% OPT	**25**	0	0	1.1664	TM 40%
10	0	0	0.8285	BMA VOL	**26**	0	0	1.2136	TM 10%
11	1	1	0.8537	NAGARCH	**27**	0	0	1.2316	TM 50%
12	1	1	0.8629	NGARCH	**28**	0	0	1.2660	TM 30% VOL
13	0	0	0.8804	TM 30% OPT	**29**	0	0	1.4144	TM 40% VOL
14	0	0	0.8842	TM 20% OPT	**30**	1	1	1.8426	TAGRCH
15	0	0	0.8941	TM 40% OPT	**31**	0	0	2.0746	TM 50% VOL
16	0	0	0.8969	TM 20% VOL	**32**	1	1	3.6332	GJRGARCH

Table 3.11.: Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity from 60 to 160 days and Moneyness from 1.05 to 1.15.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	1	1	0.0777	GARCH	**17**	0	0	0.0870	BA OPT
2	0	0	0.0823	TM 10%	**18**	0	0	0.0870	TM 30% VOL
3	0	0	0.0835	TM 30%	**19**	1	1	0.0871	NAGARCH
4	0	0	0.0838	TM 20%	**20**	0	0	0.0872	TM 20% OPT
5	0	0	0.0838	TM 40%	**21**	0	0	0.0873	TM 40% VOL
6	0	0	0.0842	TM 50%	**22**	1	1	0.0873	AVGARCH
7	0	0	0.0850	TM 10% VOL	**23**	0	0	0.0873	TM 50% OPT VOL
8	0	0	0.0855	BA VOL	**24**	0	0	0.0874	TM 10% OPT
9	0	0	0.0855	BA	**25**	0	0	0.0874	BMA VOL
10	1	1	0.0861	NGARCH	**26**	0	0	0.0874	TM 40% OPT VOL
11	0	0	0.0864	TM 50% OPT	**27**	0	0	0.0874	TM 30% OPT VOL
12	0	0	0.0865	TM 20% VOL	**28**	0	0	0.0874	TM 50% VOL
13	0	0	0.0866	TM 40% OPT	**29**	0	0	0.0875	TM 20% OPT VOL
14	0	0	0.0867	TM 30% OPT	**30**	1	1	0.0875	TGARCH
15	0	0	0.0870	BMA	**31**	0	0	0.0875	TM 10% OPT VOL
16	0	0	0.0870	BA OPT VOL	**32**	1	1	0.1554	GJRGARCH

Table 3.12.: Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity from 60 to 160 days and Moneyness more than 1.15.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0	2.3547	TM 30% OPT	**17**	1	1	3.5430	GARCH
2	0	0	2.4257	TM 40% OPT	**18**	0	0	3.6459	BA OPT VOL
3	0	0	2.4572	TM 20% OPT	**19**	**0**	0	3.6487	BA OPT
4	0	0	2.5058	TM 50% OPT	**20**	0	0	3.6976	BMA VOL
5	0	0	2.5662	TM 20% OPT VOL	**21**	0	0	3.9055	TM 10%
6	0	0	2.6568	TM 10% OPT	**22**	0	0	4.1545	TM 10% VOL
7	0	0	2.7516	TM 30% OPT VOL	**23**	1	1	4.1607	NGARCH
8	1	1	2.7825	AVGARCH	**24**	0	0	4.2183	TM 20% VOL
9	0	0	2.8470	TM 10% OPT VOL	**25**	0	0	4.3930	TM 30%
10	0	0	2.9201	BA VOL	**26**	0	0	8.7354	TM 30% VOL
11	0	0	2.9224	BA	**27**	1	1	9.6320	TGARCH
12	0	0	2.9711	BMA	**28**	0	0	10.1715	TM 40%
13	0	0	3.0505	TM 40% OPT VOL	**29**	0	0	11.7146	TM 40% VOL
14	1	1	3.1011	NAGARCH	**30**	0	0	14.9639	TM 50%
15	0	0	3.1766	TM 50% OPT VOL	**31**	1	1	16.0366	GJRGARCH
16	0	0	3.2388	TM 20%	**32**	0	0	19.5978	TM 50% VOL

Table 3.13.: Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity more than 160 days and Moneyness less than 1.05.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	1	1	1.8997	AVGARCH	**17**	0	0	2.2443	TM 40% OPT VOL
2	0	0	1.9402	TM 20%	**18**	0	0	2.2555	TM 20% VOL
3	0	0	1.9532	BA	**19**	0	0	2.2617	BA OPT VOL
4	0	0	1.9532	BA VOL	**20**	0	0	2.2623	BA OPT
5	0	0	1.9549	TM 50% OPT	**21**	0	0	2.2922	TM 10% VOL
6	0	0	1.9878	TM 40% OPT	**22**	0	0	2.2966	TM 10% OPT VOL
7	0	0	1.9927	BMA	**23**	0	0	2.3719	TM 10%
8	0	0	2.0130	TM 30% OPT	**24**	0	0	2.3927	BMA VOL
9	0	0	2.0319	TM 20% OPT	**25**	0	0	2.5371	TM 30%
10	0	0	2.0528	TM 10% OPT	**26**	0	0	4.1967	TM 30% VOL
11	1	1	2.0593	GARCH	**27**	0	0	5.6595	TM 40%
12	0	0	2.1225	TM 20% OPT VOL	**28**	0	0	6.9165	TM 40% VOL
13	1	1	2.1409	NAGARCH	**29**	1	1	8.5076	TGARCH
14	1	1	2.1671	NGARCH	**30**	0	0	8.8036	TM 50%
15	0	0	2.1970	TM 30% OPT VOL	**31**	0	0	11.8200	TM 50% VOL
16	0	0	2.2307	TM 50% OPT VOL	**32**	1	1	11.9906	GJRGARCH

Table 3.14.: Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity more than 160 days and Moneyness from 1.05 to 1.15.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0 0.769737	TM 50% OPT VOL	17	0	0	0.950648	TM 20%	
2	1	1 0.81277	NGARCH	18	0	0	0.955925	BMA VOL	
3	0	0 0.818983	BA OPT VOL	19	0	0	0.955977	TM 10% OPT VOL	
4	0	0 0.819389	BA VOL	20	0	0	0.961388	TM 10% OPT	
5	0	0 0.828206	TM 40% OPT VOL	21	0	0	0.961594	TM 20% OPT	
6	0	0 0.856755	TM 20% OPT VOL	22	0	0	0.986855	TM 30%	
7	0	0 0.862824	TM 40% OPT	23	0	0	1.018647	TM 10%	
8	0	0 0.86999	TM 50% OPT	24	0	0	1.075772	BMA	
9	0	0 0.873108	TM 30% OPT VOL	25	1	1	1.078019	NAGARCH	
10	0	0 0.874356	TM 30% OPT	26	0	0	1.102268	TM 40%	
11	0	0 0.882	TM 20% VOL	27	0	0	1.503857	TM 30% OPT	
12	1	1 0.88382	AVGARCH	28	0	0	1.554546	TM 50%	
13	0	0 0.907	BA OPT	29	0	0	2.609864	TM 40% OPT	
14	0	0 0.907349	BA	30	1	1	3.902125	TGARCH	
15	1	1 0.921058	GARCH	31	0	0	4.128393	TM 50% OPT	
16	0	0 0.946425	TM 10% VOL	32	1	1	8.924466	GJRGARCH	

Table 3.15.: Table of the out-of-sample rankings according to $RMSE of option pricing models (selected according to the best in-sample BIC criterion of the underlying volatility) and Model Averaging methods for Maturity more than 160 days and Moneyness more than 1.15.

		Maturity (days)					
Moneyness		$< \mathbf{60}$		60 to 160		$> \mathbf{160}$	
K/S		Mean	Std Dev.	Mean	Std Dev.	Mean	Std Dev.
$< \mathbf{1.05}$	Call Price	7.41	7.27	23.2	11.88	59.31	17.24
	Observations	1054		357		184	
1.05 to 1.15	Call Price	0.59	0.7	3.95	4.02	21.54	13.19
	Observations	382		224		257	
$> \mathbf{1.15}$	Call Price	1.99	4.15	0.15	0.07	3.44	2.28
	Observations	5		9		28	

Table 3.16.: Option prices descriptive statistics.

GARCH (3,2)			**AVGARCH (4,2)**			**APGARCH (1,1)**	
	MLE	MCMC		MLE	MCMC		MLE	MCMC
ω	0.0000	0.0000	ω	0.0001	0.0001	ω	0.0000	0.0001
α_1	0.0418	0.0400	α_1	0.0864	0.0810	α_1	0.0000	0.0000
α_2	0.1118	0.0981	α_2	0.0732	0.0646	γ	0.1325	0.1168
α_3	0.0316	0.0321	α_3	0.0000	0.0016	β_1	0.9337	0.9373
β_1	0.0365	0.0060	α_4	0.0000	0.0000	δ	1.2719	1.1403
	0.7634	0.8052	β_1	0.3334	0.4383			
			β_2	0.5070	0.4145			
			c	0.0067	0.0062			
	NGARCH (3,4)		**NAGARCH (2,2)**			**GJRGARCH (2,2)**	
	MLE	MCMC		MLE	MCMC		MLE	MCMC
ω	0.0000	0.0000	ω	0.0000	0.0000	ω	0.0000	0.0000
α_1	0.0354	0.0329	α_1	0.0642	0.0358	α_1	0.0000	0.0000
α_2	0.1035	0.1086	α_2	0.0642	0.0890	α_2	0.0000	0.0000
α_3	0.0000	0.0308	β_1	0.3146	0.3080	γ	0.1490	0.1692
β_1	0.2480	0.0000	β_2	0.5179	0.5424	β_1	0.8486	0.8482
β_2	0.2848	0.2612	**c**	0.0061	2.2388	β_2	0.0688	0.0609
β_3	0.0097	0.0700						
β_4	0.2853	0.4530						
δ	2.3506	2.3085						

Table 3.17.: Table of the MLE and Bayesian based estimated model parameters (an indicative example).

Date	Number of Options	Date	Number of Options
Aug-06	51	Feb-07	56
Sep-06	57	Feb-07	54
Sep-06	47	Feb-07	51
Sep-06	52	Feb-07	54
Sep-06	49	Mar-07	86
Oct-06	56	Mar-07	86
Oct-06	51	Mar-07	71
Oct-06	45	Mar-07	58
Oct-06	44	Mar-07	67
Nov-06	58	Apr-07	56
Nov-06	52	Apr-07	62
Nov-06	44	Apr-07	57
Nov-06	45	Apr-07	58
Nov-06	57	May-07	59
Dec-06	49	May-07	46
Dec-06	46	May-07	46
Dec-06	51	May-07	61
Dec-06	58	May-07	58
Jan-07	75	Jun-07	76
Jan-07	63	Jun-07	51
Jan-07	52	Jun-07	65
Jan-07	50	Jun-07	78

Table 3.18.: Number of options on the S&P500 index used to calculate each surface.

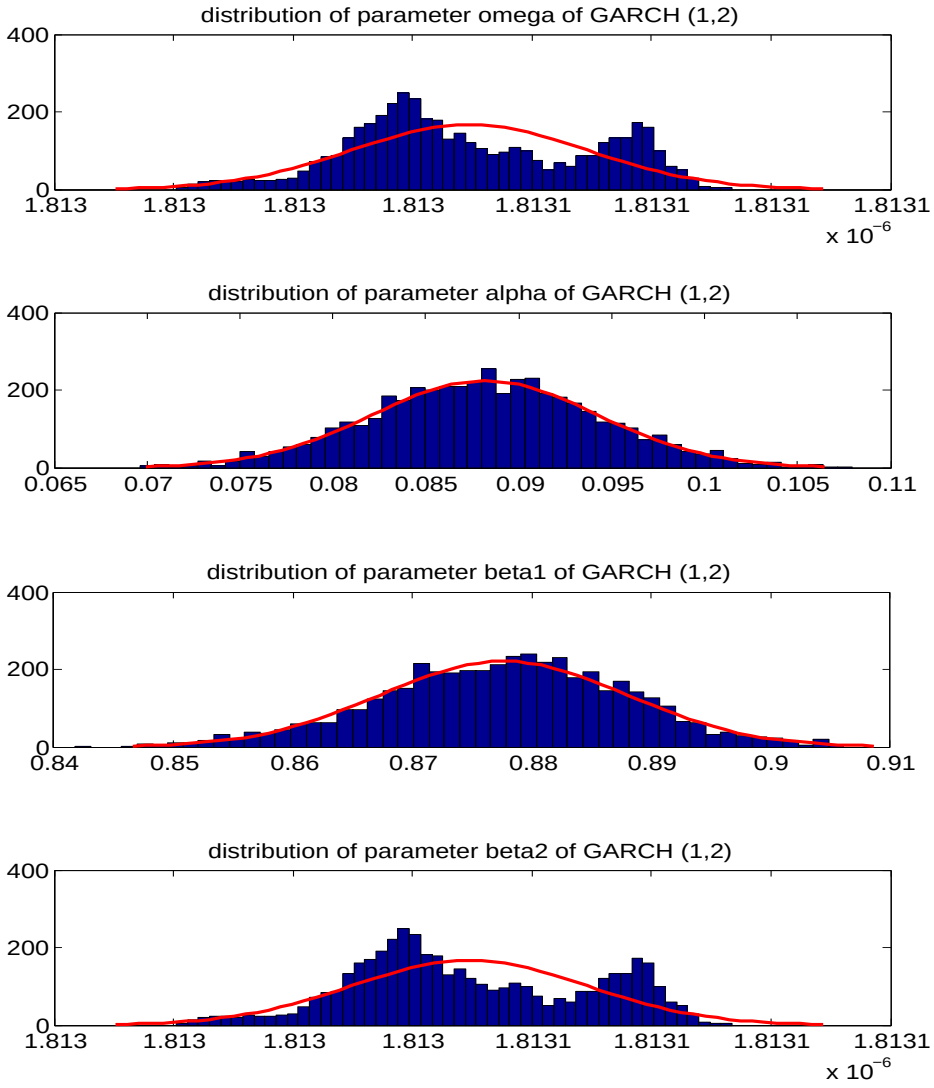


Figure 3.1.: Plot of the fitted distribution of GARCH(1,2) parameters. This is the actual distribution of each model parameter of the GARCH(1,2) along with the Normal fitted proposal distribution. The ARCH and the first GARCH parameter clearly resemble the proposal Normal . The same however does not hold here for the rest of the parameters.

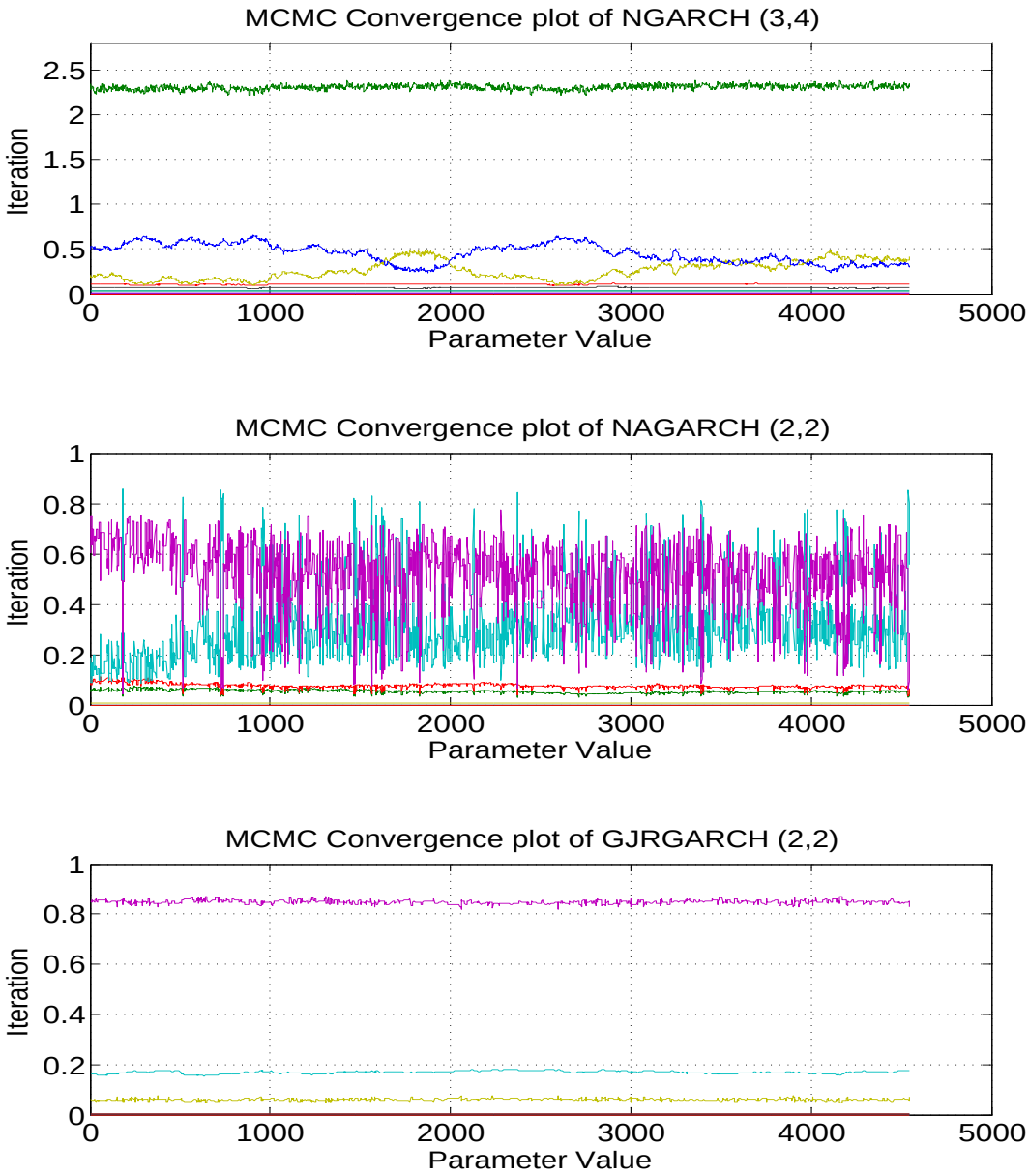


Figure 3.2.: Convergence Plots of parameters to their theoretical distribution using MCMC. This is the process of convergence for all parameters of NGARCH(3,4), NAGARCH(2,2) and GJR-GARCH(2,2) to their actual distribution using MCMC with every line standing for a different parameter.

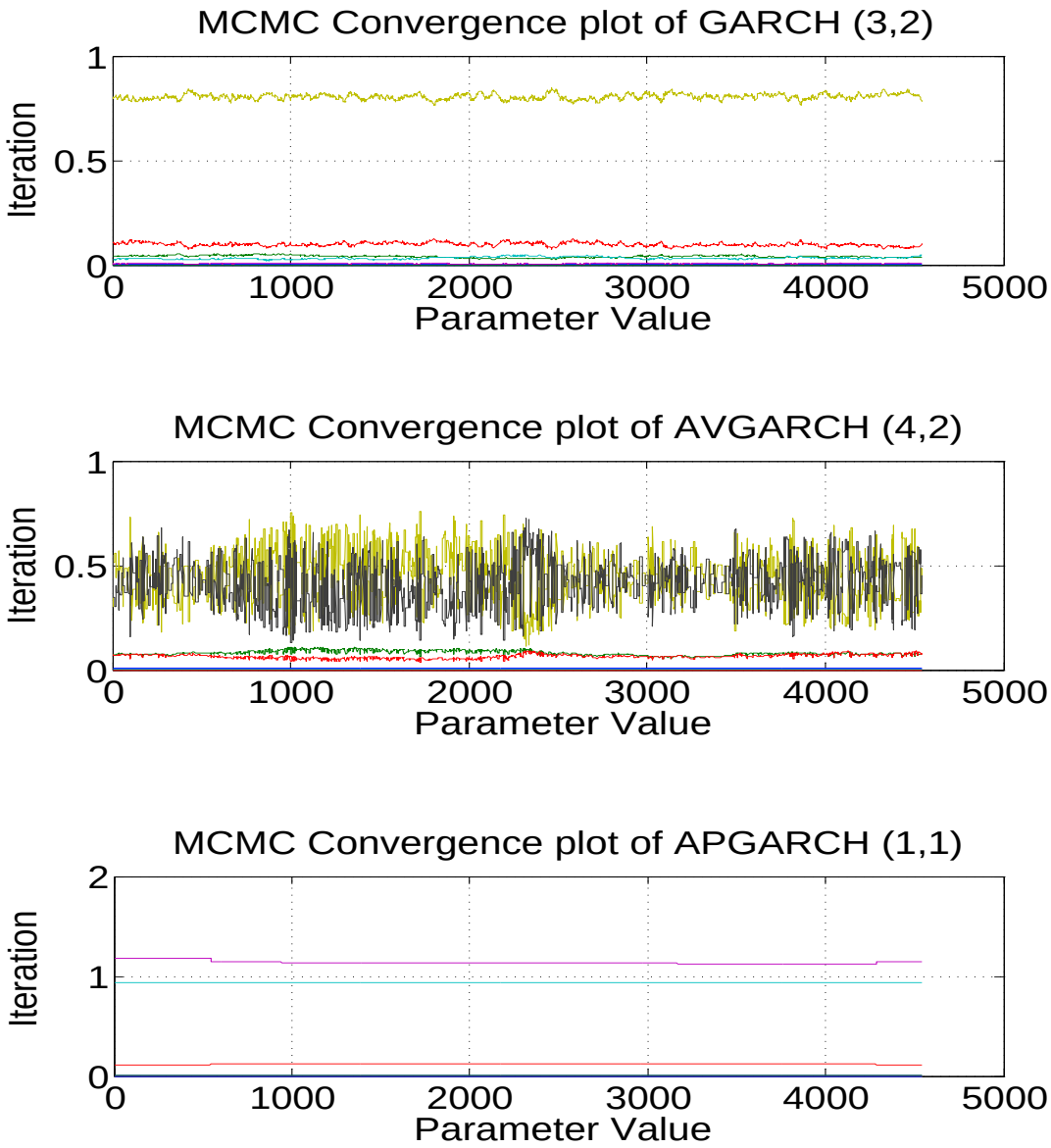


Figure 3.3.: Convergence Plots of parameters to their theoretical distribution using MCMC. This is the process of convergence for all parameters of GARCH(3,2), AVGARCH(4,2), APGARCH(1,1).

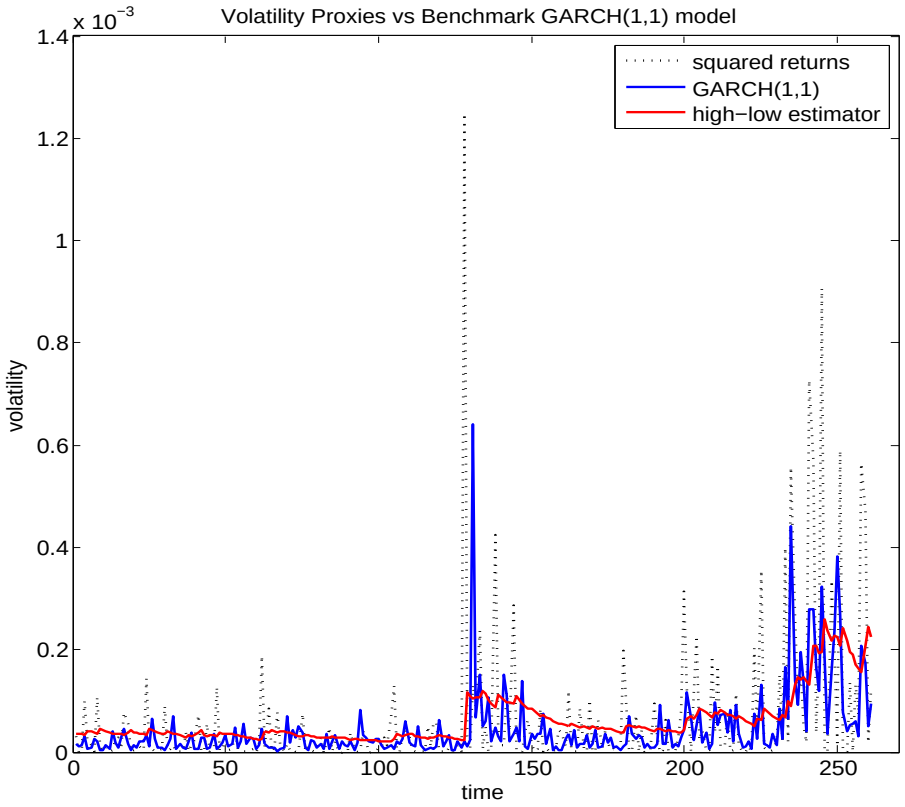


Figure 3.4.: Plot of two proxies, squared returns and high-low estimator and GARCH(1,1) for the out-of-sample period (260 days or else one year from 31/8/2006 to 1/9/2007).

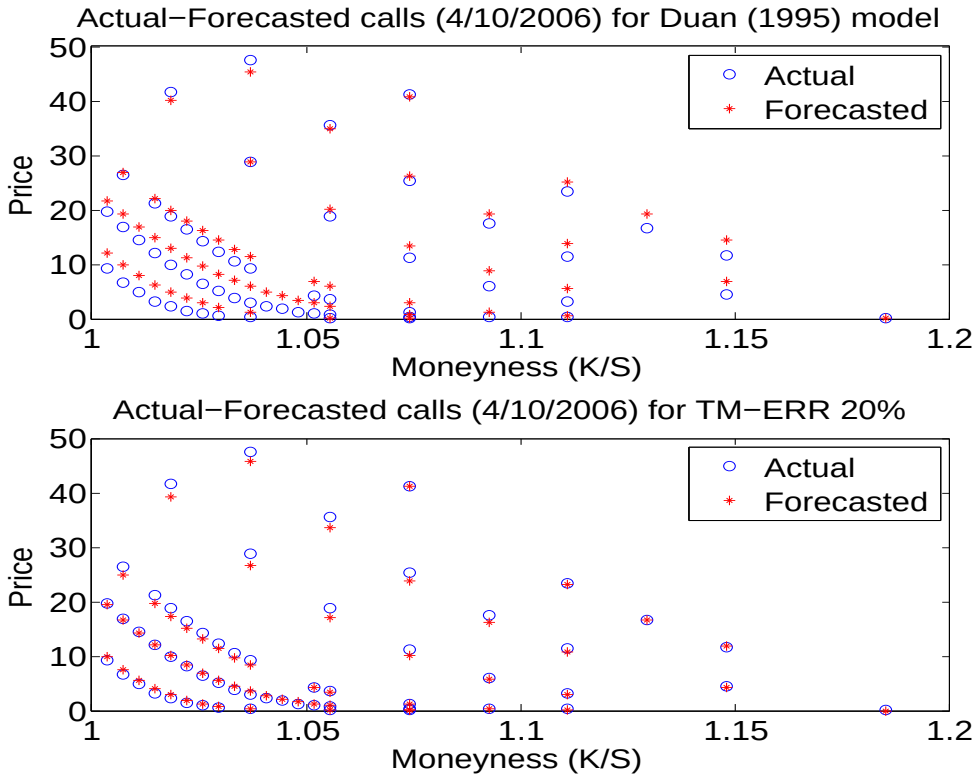


Figure 3.5.: Plot of the out-of-sample performance of Duan's option pricing model against and Thick Modelling with weights based on option $RMSE and 20% retention percentage.

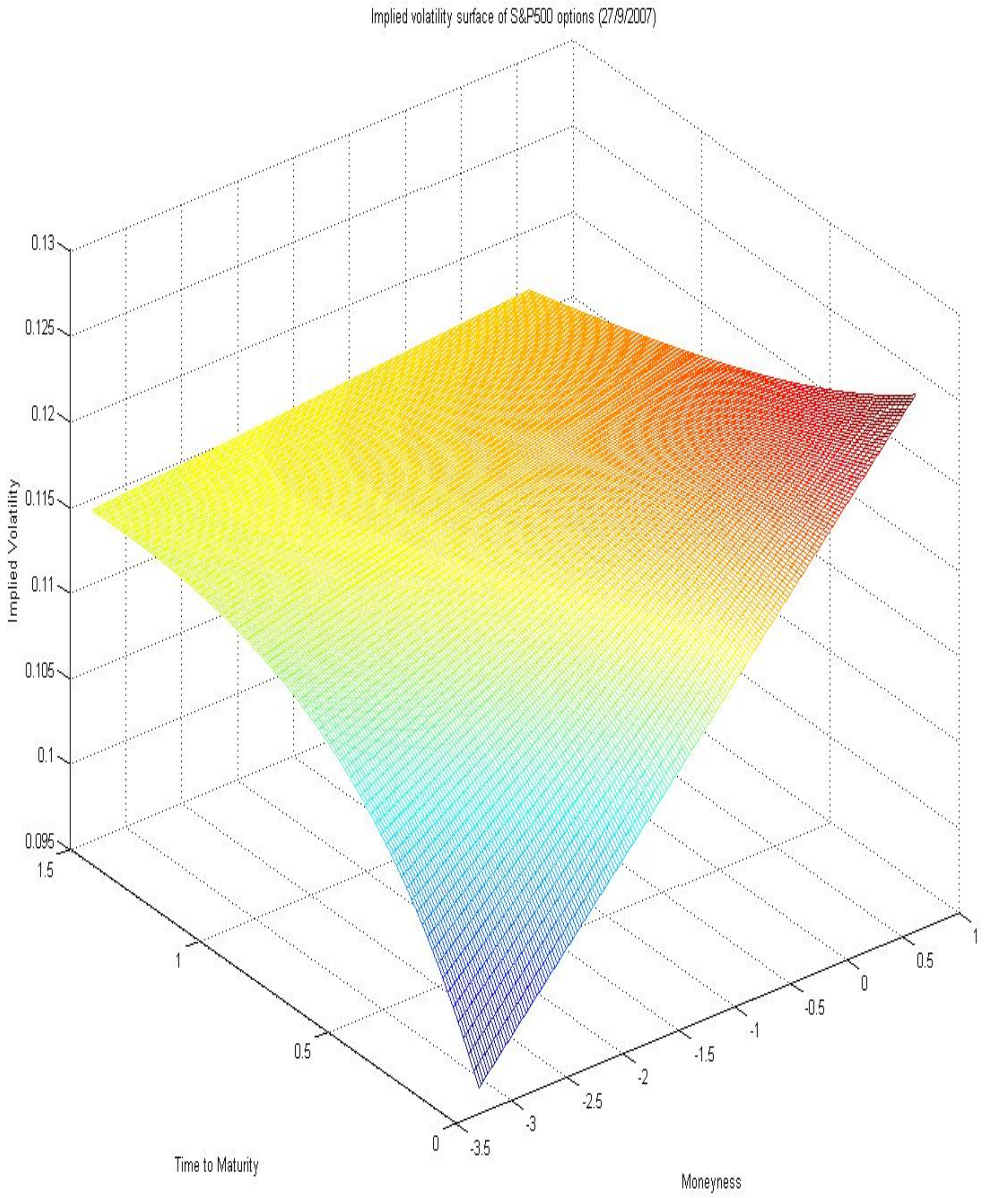


Figure 3.6.: Implied volatility surface based on options on the S&P500 index.

4. Volatility Density Forecasting and Model Averaging

4.1. Introduction

Focusing merely on the estimation of the first two moments of the probability density, while it has attracted the overwhelming majority of interest in the past, poses a major risk. This is to disregard the need to deviate from the log-normality assumption of the asset-price distribution. If we allow more attributes to play a role in the predictive distribution, then directly forecasting that density and not characteristics of it, as mean and variance themselves can suggest, becomes the imminent task to perform. One cannot help but wonder, however, why density forecasting should be the focal point, and why the importance of it is more evident now than ever before. The recent introduction of complicated structured derivative products, along with alterations in the frictionless market environment (a sufficient assumption until recently), has shaped the decision-making necessities beyond recognition. Never before has strategic risk management relied so extensively on the use of derivatives, which in turn dictated the need to make correct and not approximate inferences of the underlying asset's distribution. Cutting-edge portfolio optimisation strategies not only pointed towards density forecasting, but also rendered point or interval estimation an inadequate tool for those responsible for decision making.

However, focusing however on volatility density forecasting generates additional interest. This is because when the predictive density under scrutiny is in particular the volatility distribution, we have to remember the fact that volatility itself is an unobservable process. Inferences can be made, but validation of their accuracy is not only difficult to perform but also biased towards the variable used to approximate volatility (most commonly used proxies are the squared returns and intraday data).

From all the above we can identify in the relevant literature two distinct schools of thought addressing the problem of volatility density forecasting. One advocates the inference of the *Implied Volatility Risk-Neutral Density* from option prices when the underlying asset is actually the Implied Volatility Index. This method celebrates a wider acceptance due to Breeden and Litzenberger (1978), who established a relationship between the second derivative of the option price with respect to the strike price and the risk-neutral density. Their proposition is based on no particular distributional assumption of the underlying asset, but instead their formula is very flexible, allowing a variety of option model alternatives to be used from the researcher. The only prerequisite is a sufficient number of option data across different strike prices, from which inferences of the underlying asset can be made. This fact along with the constantly increasing interest in options on the VIX index (Implied Volatility Index of the S&P500 equity Index) justifies the popularity of the method.

At this point it is relevant to refer to some facts about the VIX Chicago Board Options Exchange (CBOE) index which will be used in the analysis. The index itself was launched in 1993 by CBOE. Originally it was used to to gauge the 30-day volatility implied by at-the-money options on the S&P 100 index. In 2003, ten years after the index was first introduced to the market, CBOE together with Goldman Sachs updated VIX. The "new" VIX is based now on the S&P 500 index, while for its estimation a wider range of call and put prices is used this time. However, it is still quoted on an annual basis. In particular to arrive at the VIX value, a range of in-the-money to out-of-the-money call options and put options of two expiration months bracketing the nearest 30-day period are selected. The implied volatility of all options of each of the selected months are estimated on a price weighted average basis in order to arrive at a single average implied volatility value for each month. Finally, results of the two months are interpolated using a 30 days constant and then a percentage derived from the square root of that result.

In particular the mathematical formula used to estimate the VIX is

$$\sigma^2 = \frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{RT} Q\left(K_i\right) - \frac{1}{T}\left[\frac{F}{K_0} - 1\right]^2, \tag{4.1}$$

where

- $$\sigma = \frac{VIX}{100} \Rightarrow VIX = 100\sigma, \tag{4.2}$$

- T is time to expiration, and,

$$T = \{M_{current\ day} + M_{settlement\ day} + M_{other\ days}\}, \tag{4.3}$$

and

$M_{\text{current day}}$ = minutes remaining until midnight of the current day,

$M_{\text{settlement day}}$ = minutes from midnight until the open time of trading (e.g. 8:30 a.m. Chicago time) on SPX settlement day,

$M_{\text{other days}}$ = total minutes in the days between current day and settlement day,

- F is the forward index level derived from option prices,

- K_0 is the first strike below the forward index level, F,

- K_i is the strike price of ith out-of-the-money options (a call if $K_i > K_0$ and a put if $K_i < K_0$, both put and call if $K_i = K_0$),

- ΔK_i is the interval between strike prices - half the difference between the strike on either side of K_i,

$$\Delta K_i = \frac{K_{i+1} - K_i}{2}. \tag{4.4}$$

Regarding the lowest strike, ΔK_i equals the difference between the lowest strike and the next higher strike. Additionally, for the highest strike, ΔK_i equals the difference between the highest strike and the next lower strike.

- R is the risk free rate to expiration, and is a bond equivalent yield of the U.S. T-bill maturing closest to the expiration dates of relevant SPX options,

- $Q(K_i)$ is the midpoint of the bid-ask spread for each option with strike K_i.

Futures on VIX were not introduced before 2004 while it took CBOE another two years, in 2006, to issue volatility options. Since then, however, it attracts a considerable interest, as it is regarded as an indicator of the future movement of the S&P500 index itself. Periods of high anticipated volatility can trigger an increase in the option price, since this could imply abrupt forthcoming moves in the equity price index. It is also called the "fear index", and when S&P500 is divided with VIX a gauge of the investor's confidence for the movement of the index can be produced. Of course a small VIX value implies low volatility, thus a high ratio and a sense of market stability. The opposite holds also. Apart from VIX, which is related to the S&P500 30-day implied volatility, there is the VXD for Dow Jones and VXN for Nasdaq100.[1]

Since option prices can reveal only density estimates in the risk-neutral world, transition from a Q-measure density forecast to the real-world equivalent predictive distribution is essential. However, calibrating the risk premium has been a subject of interest for various researchers, an issue that will also be addressed later in this chapter. A$\ddot{i}$t-Sahalia and Lo, (2000) calibrate the risk premium using an estimator related to the investors "Marginal Rate of Substitution" or the rate of the State Price Density over the real-world Density, f^Q/f^P. Starting with Merton's (1971) proposition that the ultimate objective of the investor is maximizing his wealth subject to some constraints (positive wealth etc.), the authors provide the necessary steps in order to make inferences for the Relative Risk Aversion based on the investor's risk preferences as these are expressed through his selected utility function.

The method is also discussed by Jackwerth and Carsten, (2000), Bliss and Panigirtzoglou (2001), Taylor (2005) etc. (to name a few others), while it has received a wider acceptance during the years from researchers due to its tractability. It follows naturally that by identifying any two of the three, RRA, f^Q or f^P, we are able to make inferences of the third (Bliss and Panigirtzoglou (2004)).

The second approach that will be discussed in this chapter involves using directly real-world prices (any index's prices suffice, but here the S&P500 index is used) and by applying relevant bootstrapping methods we sample

[1]For further details about the VIX or relevant indexes refer to CBOE site: www.cboe.com/VIX

from the desired distribution (it could be the empirical or any selected candidate distribution) volatility estimates for a particular point in time. However, in this case, as opposed to the previous approach where we target implied volatility distribution estimates, the objective is to forecast real-world volatility densities. For this, conditional-volatility models are required, and in particular the GARCH family models are used. This is due to the fact that no other modelling alternative has addressed the problem of volatility forecasting more efficiently than GARCH-type models have done. Since the introduction of Engle's ARCH model in 1982 GARCH family models still enjoy the same level of popularity among researchers due to their efficiency to capture some of the most well-known features of volatility (clustering, leverage effect etc.). So the modelling direction has been easy to be selected. Still though the question which one of all those GARCH type candidate alternatives is optimal to use remains. The same holds of course when the risk-neutral volatility density is under scrutiny. The number of competing option pricing schemes in the literature is vast to say the least and different specifications of each modelling approach capture different attributes of the underlying.

Naturally the question that has to be posed to the researcher is how devastating a wrong model selection can be to the final result. The solution to the problem can be traced in the use of averaging techniques. Although the idea dates far in the past, it has only recently received a revived interest due to computational advances. The thought of using approximation methods (Monte Carlo) under one model specification for option pricing was enough to daunt the common practitioner, let alone average a dozen of such models to produce an averaged scheme. Nowadays, however, the perception of computational power has been transformed beyond recognition and what seemed impossible a few decades ago is common practice today. Avoiding model misspecification risk by all means is worth the extra computational burden. On top of that, empirical evidence suggests that a better forecasting performance can be achieved when averaging is applied than one based on the best single model, as this is defined in-sample using various model selection criteria. The reason lies in the inadequacy of these criteria to identify that candidate predictive method that could be as successful in-sample as out-of-sample.

Here two of the most popular averaging processes will be applied. These

are, Thick Model Averaging (Granger and Jeon, 2004) and Bayesian Approximation. When the risk-neutral volatility density forecasting is under scrutiny, averaging techniques will be applied using the option-pricing models and then Breeden and Lichtenberger's method for deriving the density of the underlying based on option prices. When the real-world volatility density is under scrutiny the different GARCH type models for conditional-volatility will be averaged directly in order to produce desired density forecasts.

4.2. Literature Review

In general there are two approaches governing density forecasting. The first relates to the data as these are observed in the market, so for that reason it focuses directly on the so called "real-world" density. The second approach, however, and the most favourable of the two, is based on the observed option prices related to the particular asset of interest, and from these prices inferences of the asset's risk-neutral distribution can be made instead.

4.2.1. Risk-Neutral Density Forecasting

Ross (1976) proves that one can recover the risk-neutral distribution from a continuum of option prices, while Breeden and Litzenberger (1978)[2] establish in particular a relationship to derive this distribution, connecting the second derivative of the option price with respect to the strike price with the risk-neutral density. To be more precise, this relationship can be formulated as follows.

Let us define first S_t as the value of an asset at time t, with $c(S_T)$ being the payoff at time T $(T > t)$ of a call option written on that asset which has a risk-neutral density $f_t^Q(S_T)$. According to Ross (1976) and Cox and Ross (1976), the call price equals (an equivalent expression can be postulated for put options as well)

[2] A lot of authors reference as well Ross (1976) and Banz and Miller (1978) work on the area of deriving the risk-neutral density from observed option prices.

$$
\begin{aligned}
c\left(S_t, K, T\right) &= e^{-rT} \int_0^{\infty} \left(S_T - K\right)^+ f_t^Q\left(S_T\right) dS_T \\
&= e^{-rT} \int_K^{\infty} \left(S_T - K\right) f_t^Q\left(S_T\right) dS_T \\
&= e^{-rT} \int_K^{\infty} S_T f_t^Q\left(S_T\right) dS_T - e^{-rT} K p^Q, \qquad (4.5)
\end{aligned}
$$

where r is the risk-free rate and K the strike price and p^Q the probability of exercising the call under the Q-measure.

Taking under consideration the above Breeden and Litzenberger, prove in turn that the risk-neutral density can be connected with the pricing equation via the following formula:

$$
f_t^Q\left(S_T\right) = e^{rT} \left. \frac{\partial^2 c\left(S_t, K, T\right)}{\partial K^2} \right|_{K=S_T}. \qquad (4.6)
$$

This is the approach followed here which relates also to the formation of a butterfly spread. In other words, a butterfly spread can be used in order to approximate the risk-neutral density. To construct that spread we first assume a portfolio that consists of a long position on two calls options, one with strike $K - s$ and another with strike $K + s$, and prices $c\left(S_t, K - s, T\right)$ and $c\left(S_t, K + s, T\right)$ respectively. Also in this portfolio we take a short position on two identical call options with strike K and price $c\left(S_t, K, T\right)$. In the end we should have created the desired butterfly spread with total price $P\left(S_t, K, T\right)$ given by:

$$
P\left(S_t, K, T, s\right) = \frac{c\left(S_t, K - s, T\right) - 2c\left(S_t, K, T\right) + c\left(S_t, K + s, T\right)}{s^2}, \qquad (4.7)
$$

where $\frac{1}{s^2}$ are the units of the option. At time T its payoff is,

$$
P\left(S_T, K, T, s\right) = \frac{s + \left|S_T - K\right|}{s^2} \mathbf{1}_{S_T \in [K-s, K+s]}. \qquad (4.8)
$$

Now as s approaches zero the payoff approximates

$$\lim_{s\to 0} P\left(S_T, K, T, s\right) = \begin{cases} \infty & \text{if } S_T = K, \\ 0 & S_T \neq K. \end{cases} \tag{4.9}$$

If we assume that at time t the payoff of the option equals the discounted expected risk-neutral value under the Q-measure, then the above limit, $\lim_{s\to 0} P\left(S_T, K, T, s\right)$, connects the risk-neutral density with the terminal option value:

$$\lim_{s\to 0} P\left(S_T, K, T, s\right) = e^{-rT} E_t^Q [\lim_{s\to 0} P\left(S_T, K, T, s\right)] = e^{-rT} f_t^Q\left(S_T\right). \tag{4.10}$$

Rewriting (4.3), we get

$$P\left(S_t, K, T, s\right) = [\frac{c\left(S_t, K-s, T\right) - c\left(S_t, K, T\right)}{s} - \frac{c\left(S_t, K, T\right) - c\left(S_t, K+s, T\right)}{s}]/s \tag{4.11}$$

which implies that

$$\lim_{s\to 0} P\left(S_T, K, T, s\right) = \left.\frac{\partial^2 c\left(S_t, K, T\right)}{\partial K^2}\right|_{K=S_T}. \tag{4.12}$$

From (4.6) and (4.8), finally we get

$$\begin{aligned} f_t^Q\left(S_T\right) &= e^{rT} \frac{\partial P\left(S_t, K, T, t\right)}{\partial K^2} \approx e^{rT} \frac{\Delta P\left(S_t, K, T\right)}{\left(\Delta K\right)^2} \\ &\approx e^{rT} \frac{-2c\left(S_t, K, T\right) + c\left(S_t, K-s, T, t\right) + c\left(S_t, K+s, T\right)}{s^2} \end{aligned} \tag{4.13}$$

where S_t is the spot price at time t, K the strike price, s is a very small amount used to separate contingent claims on the same asset with different strikes (all centered around K) and T is the time to maturity. If we had $1/s$ units of the butterfly spread, then we can say that as s approaches zero the butterfly spread approximates an Arrow-Debreu security (namely also state price and in accordance the corresponding risk-neutral density is called state price density[3]) that pays

[3] A good reference including some of these introductory to the financial and option pricing definitions is: Bahra (1997)

$$\lim_{s \to 0} P\left(S_T, K, T, s\right) = \begin{cases} 1 & \text{if } S_T = K, \\ 0 & S_T \neq K. \end{cases} \tag{4.14}$$

In order to estimate the risk-neutral density given a collection of option prices on the same underlying asset with different strikes but with the same maturity, we can identify mainly two schools of thought. The first relates to parametric while the other to non-parametric estimation methods. Different categorizations do exist, but here we employ the most common practice of dividing the vast literature related to the subject according to those two wider predetermined categories. More extensive literature reviews on the subject of recovering the risk-neutral distribution are provided by Bahra (1997), Jackwerth (1999), and more recently Bliss and Panigirtzoglou (2002). Here we will loosely follow the categorization as presented in Jackwerth.

According to this the parametric methods can be subdivided into 3 further categories.

1. Risk-neutral density related models

   - Expansion methods

     Under this spectrum Jarrow and Rudd (1982) suggest deviating from the lognormal density assumption on the underlying by adding to the polynomials alleged distribution . These account for a dependency structure which incorporates higher moments such that more complicated attributes of the asset can be captured. Similarly Edgeworth expansion is used by Longstaff (1995, normal deviate related expansion), Rubinstein (1998, binomial density expansion) and more recently Giamouridis *et al.* (2002, 2005). In contrast with Edgeworth expansion that uses cumulants, Gram-Charlier expansion uses moments of the distribution to deviate from lognormality and its application in the risk-neutral density estimation is introduced by Corrado and Su (1997).

   - Generalized distribution methods

     Here more general distributions are employed, such as Weibull, Gamma (Savickas (2005)), Generalized Beta (McDonald *et al.*

(1987, 1995)), GEV (Markose and Alentorn (2005)), as well as many other alternatives including Generalized Lambda (Corrad, 2001), Burr XII distribution (Sherric *et al.* (1992)) etc.

- Mixtures of distributions

  The most common encounter in this category is a mixture of two lognormals. It could also possibly be a mixture of three or even more distributions, although the unavoidable explosion of parameters restricts the utilization of this approach. Usually this method can be associated with the preconception that a particular number of states of the world (two, three etc.) can only exist in the future, each one with probability p_i where $\sum_{i=1}^{s} p_i = 1$ and s is the number of states. Authors that have applied this method include among others Ritchey (1990, mixture of normals), Bahra (1997) and Melick and Thomas (1997, mixture of three lognormal), Soderlind and Svensson (1997), Abadir and Rockinger (2003), Giacomini *et al.* (2008).

2. Implied-volatility (or volatility-smile) models

   Under this approach, from the observed option prices the implied volatility is derived and then fitted to a lower degree polynomial which represents the market implied-volatility function. This function can later be used to derive option prices, which in turn can be translated into probabilities under the Breeden's and Litzenberger's transformation. First Shimko (1993) proposes the method which is later applied by Malz (1997a, 1997b), Dumas *et al.* (1998). This approach, however, introduces higher estimation bias than that which uses directly observed option prices to derive the risk-neutral density.

3. Stochastic-process models

   Here particular stochastic processes may be used to model the dynamics of the underlying asset of interest. This approach has the additional advantaged that a more complicated process for the underlying can be assumed which could actually incorporate a series of irregularities of the variable as these are observed in the market (occurrence of jumps, mean reversion etc.). Undeniable great contributors to this category with models describing a stochastic dependence structure of

the volatility of the underlying are Hull and White (1988), Heston (1993), more recently Bates (1996) and many others.

In addition the nonparametric methods can be divided into the following subgroups:

1. Implied trees

   Rubinstein (1994) uses implied binomial trees to make inferences about the risk-neutral density of the underlying. The posterior end node probabilities are derived by minimizing the sum of squared deviations error (we calculate that using the prices estimated when probabilities are assigned to the tree nodes and the observed option prices). Jackwerth and Rubinstein (1996) suggest the use of different target functions to solve the minimization problem.

2. Maximum-entropy methods

   Under this approach inferences are made using only observed option prices while the method produces a competent density estimate even when limited observations are available. Here the RND is approximated using the maximum entropy distribution, which is derived by maximizing Shanon's entropy measure[4] conditioning on moment restrictions to ensure positivity of the outcome (see Buchen and Kelly (1996), Stutzer (1996), Hawkins (1997)).

3. Curve fitting methods

   - Kernel methods

     Aït-Sahalia and Lo (1998, 2000) and Aït-Sahalia *et al.* (2001) suggest using a non-parametric option pricing formula, twice differentiable with respect to strike, to make estimates of the underling's distribution. The value of the non-parametric estimator is calculated by fitting observed option prices by a kernel regression . Afterwards this estimator can be used to derive the risk-neutral density following Breeden and Litzenberger. The ultimate objective of this method is to reduce the dimension of the parameter vector related to the estimation problem, while at the same time provide a trustworthy estimator.

[4]For further details regarding the method refer to the original paper: Jaynes (1957)

- Regularization methods

  Here we can identify those methods that use the option prices directly to make inferences for the RND and the ones that use the implied volatility instead, and fit smoothing splines to make relevant inferences, e.g. Campa *et al.* (1998) and Bliss and Panigirtzoglou (2002).

A concluding remark is the fact that we should not forget that adjustments to the option-pricing equation in order to incorporate the price of risk are required. This is very important in order to be able to price under the P-measure (real-world measure). Options only reveal information associated with a risk-neutral environment, and transitioning to a non-martingale reality is the ultimate objective (although the calibration of this risk-premium parameter will be discussed at a later point).

4.2.2. Stochastic-Volatility Option Pricing

Having mentioned all the above, Breeden and Litzenberger's method that can lead to inferences about the distributional patterns of the underlying actually makes no assumption regarding the modelling requirements of the contingent claim. One should though have in mind above all that the principal focus of this research is volatility density forecasting. Therefore further consideration is needed related to the particular forces that drive volatility derivative products, as these can reveal the implied volatility distribution. In order to capture these particular dynamics (stochastic pattern of the volatility, central tendency, correlated jumps etc.), a wide collection of modelling alternatives has been used all these years. Relative research regarding volatility derivative pricing goes back in the early nineties. Of the first to be involved with predicting the volatility density are Grünbichler and Longstaff (1995). They provide an analytical formula to very conveniently track volatility future and option prices. Their work could only offer a glimpse of the interest that has sparked in the years that have followed as regards to the significance of volatility derivatives.

Grünbichler and Longstaff assume a square root process for the underlying, as follows:

$$dV_T = (\alpha - \kappa V_T)\, dt + \sigma \sqrt{V_T} dZ_T. \tag{4.15}$$

Based on Feller (1951) and Cox and Ross (1985), they proved that the risk-adjusted underlying asset is noncentrally $\chi^2(\nu, \lambda)$ distributed. This gives the analytical solution to the PDE of the option pricing equation as follows:

$$c(V_T, K) = D_T E^Q \left[(V_T - K)^+ \right], \tag{4.16}$$

where $D_T = e^{-rT}$ is the discount factor (r is the riskless rate), V is the underlying volatility index, K is the strike price, T is the time to maturity.

Progress however has been relatively slow, and only recently interest has begun to rekindle when options on VIX started to be traded, putting the valuation of volatility derivatives back in the limelight. In general VIX gauges the implied volatility of the S&P 500 index over a 30-day period and is reported on an annual base. Although it started being traded in 1993, there is an apparent lag in the introduction of contingent claims on the index, since it is not before 2004 that futures on VIX made their first appearance, and it took another couple of years (2006) for options on the same index to be introduced.

From an academic and investment point of view, attention was drawn instantly. Models to evaluate option prices on the index varied from those including mean-reverting processes, stochastic-volatility processes, those that incorporated jumps in conjunction with mean reversion or/and stochastic-volatility, double mean-reverting stochastic-volatility models (Gatheral (2008), Li (2009)) and many other propositions. Of the first to introduce pricing formulae are Detemple and Osakwe (2000). Their suggestions are based on well-established asset-pricing diffusion processes which are now being utilized for option volatility modelling purposes. These propositions include a Geometric Brownian Motion process (Merton (1973)), a Mean-Reverting Gaussian process (Vasicek (1977)), a Mean-Reverting Square-Root process (Cox, Ingersoll and Ross (1985)) and finally a log-Mean-Reverting Gaussian process. In particular the authors suggest that the volatility can be described using one of the following equations:

Model	*SDE*	*Solution*
GBM	$dV_T = V_T\left[\left(2\mu+\sigma^2\right)dt + 2\sigma dZ_T\right]$	$V_T = V_0\exp\left(\left(2\mu+\sigma^2\right)t + 2\sigma dZ_t\right),$ $Z_t \sim N(0,t)$
MRGP	$dV_T = (\alpha - \lambda V_T)\,dt + \sigma dZ_T$	$V_T = V_0\phi_t + A_t + \alpha_t w_t,$ $w_t \sim N(0,1)$ $\phi_t = e^{-\lambda t}$ $A_t = \frac{\alpha}{\lambda}(1-\phi_t)$ $a_t = \frac{\sigma}{\sqrt{2\lambda}}\sqrt{(1-\phi_t^2)}$ $w_t \equiv \frac{1}{a_t}\left[\int_0^t e^{-\lambda(t-s)}dZ_s\right]$
MRSRP	$dV_T = \left(\sigma^2 - 2\lambda V_T\right)dt + 2\sigma\sqrt{V_T}dZ_T$	$V_t = \left(\sqrt{V_0}\phi_t + a_t w_t\right)^2,$ $w_t \sim N(0,1)$
LMRGP	$d\ln V_T = (\alpha - \lambda\ln(V_T))\,dt + \sigma dZ_T$ or $dV_T = V_T\left(\alpha + \frac{1}{2}\sigma^2 - \lambda\ln(V_T)\right)dt + V_T\sigma dZ_T$	$V_T = V_0^{\phi_t}e^{A_t+\alpha_t w_t},$ $w_t \sim N(0,1)$

Table 4.1.: Assumptions of the volatility process according to Detemple and Osakwe (2000).

An important aspect, though, that is absent in the two previous scientific papers (Grünbichler and Longstaff, Detemple and Osakwe) is the fact that from time to time jumps in the volatility price are observed. One of the

first to discuss this attribute is Merton (1976), who extends the Black and Sholes model in order to incorporate this discrete-time feature (price jumps do not happen constantly). According to Merton the underlying asset can be described by the following equation:

$$dS_t = \mu S_t dt + \sigma S_t dZ_t + \kappa S_t dQ_t, \tag{4.17}$$

where κ is the jump size, and it holds that

$$\ln(1+\kappa) \sim N\left(\ln(1+\bar{\kappa}) - \frac{1}{2}\gamma^2, \gamma^2\right),$$

with $\bar{\kappa}$ the average jump size and γ^2 the volatility of the proportional jump size. Merton's model can be solved analytically. Just extending the Black and Scholes equation and incorporating jumps could not suffice, however, when volatility options are under scrutiny. Adding to the process a mean-reversion component seemed the next more apparent extension. Clewlow and Strickland (2000), opting for a better model to describe energy derivatives, have noticed first that the assumption of a constant volatility process as this is implied by Black Sholes and Merton is not appropriate, and second that electricity prices do convert to a long-term mean despite the periodical occurrence of price jumps.

The fact that the underlying asset has time-dependent mean-reverting properties may first discussed, however, by Schwartz (1997). In his one-factor stochastic-volatility model no jump process is incorporated, however. This is an extension added by Clewlow and Strickland (2000). The model's SDE is

$$dS_t = \alpha(\mu - \ln S_t) S_t dt + \sigma S_t dZ_t, \tag{4.18}$$

where μ is the long-term mean and α is the rate of mean reversion. Accounting for risk, the equation can be adjusted

$$dS_t = \alpha(\mu - \lambda - \ln S_t) S_t dt + \sigma S_t dZ_t. \tag{4.19}$$

To model volatility derivatives the relative equation associated to the log price (accounting for the price of risk as well) is also important enough to be mentioned here:

$$d\ln S_t = \alpha\left(\mu - \lambda - \frac{\sigma^2}{2\alpha}\ln S_t\right)dt + \sigma dZ_t. \tag{4.20}$$

Stepping forward from Merton's model, however, implied sacrificing its analytical tractability. Monte Carlo simulation is the next direct choice for solving the PDE, while maximum likelihood methods can also be an option here. By adding jumps to the prices (as Clewlow and Strickland suggested) the SDE is extended as follows:

$$dS_t = \alpha\left(\mu - \ln S_t\right)S_t dt + \sigma S_t dZ_t + \kappa S_t dQ_t, \tag{4.21}$$

where $\alpha, \mu, \sigma, \kappa$ and the distribution of $\ln(1+\kappa)$ are as defined before. Readjusting in order to incorporate the market price of risk we can write

$$dS_t = \alpha\left(\mu - \lambda - \ln S_t\right)S_t dt + \sigma S_t dZ_t + \kappa S_t dQ_t. \tag{4.22}$$

Until now in the models we have referred to the parameters of the underlying asset's distribution (mean , variance) remain constant over time. First Heston (1993) adjusted the GBM model to account for time dependency in the volatility of the underlying:

$$dS_t = \mu S_t dt + \sigma_t S_t dW_{1t}. \tag{4.23}$$

Bakshi, Cao and Chen add jumps to Heston's model to account for the fact that occasionally prices exhibit sudden upward or downward moves. Additionally, influential work related to volatility modeling, such as this of Hull and White (1987), Stein and Stein (1991), Ball and Roma (1994), Fouque *et al.* (1998), Fournie *et al.* (1997), Heston (1993) etc, favours the notion of mean reversion. Thus, in an attempt to successfully capture this attribute (see also Derman *et al.* and Dupire), we construct a model were volatility is represented by an Ornstein Uhlenbeck process (at first no jumps are added to the dynamics of the underlying). The equations that describe such a model are given below:

$$\begin{aligned} dS_t &= \mu S_t dt + \sigma_t S_t dW_{1t}, & (4.24)\\ d\sigma_t^2 &= \eta\left(\vartheta - \sigma_t^2\right)dt + \gamma\sigma_t dW_{2t}, & (4.25) \end{aligned}$$

where W_{1t}, W_{2t} are ρ_{12} correlated Geometric Brownian motions with volatility being marginally Log-Normally distributed.

Another alternative in option pricing is to assume a stochastic volatility process driven by a Lévy process. The most popular in this category is the Barndorff-Nielsen and Shephard (2001) model, which assumes for the underlying a stochastic volatility that follows a non-Gaussian Ornstein-,Uhlenbeck process driven by a subordinator $(z_t)_{t\geq 0}$, $\lambda > 0$. This proposition belongs to a more general category which accounts for the fact that as prices of the underlying increase the volatility falls. This so-called leverage effect in connection with Heston's concept of a time-dependent rather that constant volatility attribute underpins the Barndorff-Nielsen and Shephard model's driving mechanism. In particular the asset price is

$$S_t = S_0 exp(r_t + X_t), \tag{4.26}$$

where X_t is described by the following SDE:

$$\begin{aligned} d\ln X_t &= \left(\mu + \beta\sigma_t^2\right)dt + \sigma_t dW_t + \rho dz_{\lambda t}, \\ d\sigma_t^2 &= -\lambda\sigma_t^2 dt + dz_{\lambda t}, \quad \sigma_0^2 \geq 0, \end{aligned} \tag{4.27}$$

where $\sigma_0^2 \geq 0$, $\rho \leq 0$, $z_t(\ t > 0)$ is the Background Driving Lévy Process (BDLP) with $z_0 = 0$ which is non-Gaussian with positive non-decreasing independent increments (i.e. a subordinator) and λ some positive constant. λ in particular plays a significant role, since it controls not only the frequency but also the rate of decay of the jump in volatility. Also W_t is independent of z_t, and ρ is the parameter that captures the asymmetric response that a change in price causes to volatility. Also the process is a martingale if

$$\begin{aligned} \beta &= -\frac{1}{2}, \\ \mu &= -\int_{\Re_+} (e^{\rho x} - 1)\nu dx, \end{aligned}$$

where ν is the coefficient that determines the size of the jump for the volatility and the price. Adjusting from the risk-neutral Q to the objective measure we get the transformed relationship for the log price:

$$d\ln X_t = \left(r - q - \lambda\nu\left(-\rho\right) - \frac{1}{2}\sigma_t^2\right)dt + \sigma_t dW_t + \rho dz_{\lambda t}. \tag{4.28}$$

Most common choices for the BNS model are when the volatility follows a Gamma or Inverse Gamma distribution. In other words in these cases the volatility of the underlying asset is Gamma or Inverse-Gamma marginally distributed and exhibits jumps. Such models have been favoured by researchers since they offer tractable characteristic functions for the underlying.

This model category is directly comparable when we impose continuous-time limits on discrete time volatility models. First Nelson (1990) suggested the following:

$$y_n = \sigma_t dB_t^{(1)},\ t \geq 0, \tag{4.29}$$

$$d\sigma_t^2 = \left(\beta - \eta\sigma_t^2\right)dt + \phi\sigma_t^2 dB_t^{(2)},\ t > 0, \tag{4.30}$$

where $B_t^{(1)},\ B_t^{(2)}$ are independent Brownian motions, $\beta > 0,\ \eta \geq 0$, and $\phi \geq 0$ are constants. The limitation of this process is that there are two independent sources of uncertainty. On the other hand in a GARCH model we find just one source of variability, the innovation process, which is modelled as a white noise. This twofold uncertainty can have complications when pricing options. Another advantage of a simple GARCH model is that large socks are interpreted immediately as large innovations thus, have a direct impact on the modelling outcome of future volatility. Unfortunately, Nelson's model does not share the same feature. Corradi (2000) tried to modify the above model in order to have just one random part. What he proposes though, reduces to a deterministic process, an undesirable property when modelling volatility. Klüppelberg, Lindner, and Maller (2004) however, introduced a complete different approach to continuous-time modelling of volatility, presenting the COGARCH model.

COGARCH(1,1) model is the continuous time analogous to the GARCH(1,1) model driven as well by one single source of random variability, which in this case however, is the background driving Lévy process (BDLP). Initially let remind us that according to Black and Scholes model (1973) the log-price process $G_t = \log S_t$ is modelled by a Geometric Brownian motion.

$$G_t = G_0 + \mu t + \sigma B_t,\ t \geq 0. \tag{4.31}$$

Given S_0 the price process is the solution to the following SDE:

$$\frac{dS_t}{S_t} = \left(\mu + \frac{\sigma^2}{2}\right) dt + \sigma dB_t,\ t > 0. \tag{4.32}$$

The log-price can also be written as

$$G_t = G_0 + \int_0^t \mu_s ds + \int_0^t \sigma_{s-} dL_s,\ t \geq 0, \tag{4.33}$$

where L is a Lévy process and $\mu,\ \sigma$ are adapted càdlàg processes. The Lévy-Itō representation states that

$$\begin{aligned} L_t = \gamma_L t + \tau_L B_t + & \int_{(0,t]\times\{|x|>1\}} x J_L\left(ds \times dx\right) + \ldots \\ & \lim_{\varepsilon \downarrow 0} \int_{(0,t]\times\{|x|>1\}} x\left(J_L\left(ds \times dx\right) - ds\Pi_L\left(dx\right)\right), \end{aligned} \tag{4.34}$$

where J_L is a Poisson random measure on $[0,t]\times R$ with intensity $dt\Pi_L\left(dx\right)$.

As a Euler approximation of the above Lévy log-price model we obtain for the martingale part

$$\begin{aligned} y_n &= \varepsilon_n \sigma_{n,disc}, & (4.35)\\ \sigma^2_{n,disc} &= \beta + \lambda y^2_{n-1} + \delta \sigma^2_{n-1,disc}\ \ n \in N, & (4.36) \end{aligned}$$

where $\sigma_{n,disc} := \sqrt{\sigma^2_{n,disc}}$ is the volatility process modelled as a random recurrence equation and $(\varepsilon_n)_{n\in N_0}$ is a sequence of iid random variables independent of $\sigma^2_{0,disc}$ with $N = \{1,2,3,\ldots\}$ and $N_0 = N \cup \{0\}$. The solution to the variance equation is as follows:

$$\begin{aligned} \sigma^2_{n,disc} &= \beta + \lambda y^2_{n-1} + \delta \sigma^2_{n-1,disc} \\ &= \beta \sum_{i=0}^{n-1} \prod_{j=i+1}^{n-1} \left(\delta + \lambda \varepsilon_j^2\right) + \sigma_0^2 \prod_{j=0}^{n-1} \left(\delta + \lambda \varepsilon_j^2\right), \end{aligned} \tag{4.37}$$

$$\sigma^2_{n,disc} \quad \xrightarrow{d} \sigma^2 \infty \stackrel{d}{=} \beta \sum_{i=0}^{\infty 1} \prod_{j=1}^{i} \left(\delta + \lambda \varepsilon_j^2\right), \tag{4.38}$$

$$\begin{aligned} \sigma^2_{n,disc} &= \left(\beta + \int_0^n \exp\left\{\sum_{j=0}^{\lfloor s \rfloor} \log\left(\delta + \lambda \varepsilon_j^2\right)\right\} ds + \sigma^2_{0,disc}\right) \ldots \\ &\quad \exp\left\{\sum_{j=0}^{n-1} \log\left(\delta + \lambda \varepsilon_j^2\right)\right\}. \end{aligned} \tag{4.39}$$

In continuous time the error term ε_n is replaced by jumps of a Lévy process. Let $(L_t)_{t\geq 0}$ be a Lévy process with jumps $\Delta L_t = L_t - L_{t-1}$ $t \geq 0$ and let $0 < \delta < 1, \lambda \geq 0$. Finally we define an auxiliary càdlàg process $(X_t)_{t\geq 0}$, where

$$X_t = -t \log \delta - \sum_{0<s\leq t} \log\left(1 + \frac{\lambda}{\delta} (\Delta L_s)^2\right),\ t \geq 0. \tag{4.40}$$

The volatility process $(\sigma_t)_{t\geq 0}$ (left-continuous) equals

$$\sigma_t^2 = \left(\beta \int_0^t e^{X_s} ds + \sigma_0^2\right) e^{-X_{t-}},\ t \geq 0, \tag{4.41}$$

where $\sigma_t := \sqrt{\sigma_t^2}$, $\beta > 0$ and σ_0^2 a finite random variable. We are now ready to define the integrated continuous time GARCH process (COGARCH) $(G_t)_{t\geq 0}$as a càdlàg process satisfying

$$G_t = \int_{(0,t]} \sigma_{t-} dL_t,\ t \geq 0, \tag{4.42}$$

or

$$dG_t = \sigma_t dL_t,\ t \geq 0,\ G_0 = 0. \tag{4.43}$$

The logarithmic price G, jumps at the same time L does, with jump size

$$\Delta G_t = \sigma_t \Delta L_t,\ t \geq 0. \tag{4.44}$$

$(X_t)_{t\geq 0}$ is spectrally negative, has drift $-\log \delta$, no Gaussian part, Lévy measure

$$\Pi_X([0,\infty)) = 0 \quad \Pi_X([-\infty,-x)) = \Pi_L\left(\left\{|y| \geq \sqrt{(e^x-1)\frac{\lambda}{\delta}}\right\}\right), \forall x > 0. \tag{4.45}$$

Finally,

$$d\sigma_t^2 = \left(\beta + \log\delta\sigma_{t-}^2\right)dt + \frac{\lambda}{\delta}\sigma_{t-}^2 d\left[L,L\right]_t^{(d)} \tag{4.46}$$

where,

$$\begin{aligned} d\left[L,L\right]_t^{(d)} &= \sum\nolimits_{0<s\leq t}\left(\Delta L_s\right)^2, & (4.47)\\ \sigma_t^2 &= \sigma_0^2 + \beta t + \log\delta\int_0^t \sigma_s^2 ds + \frac{\lambda}{\delta}\sum_{0<s\leq t}\sigma_{s-}^2\left(\Delta L_s\right)^2. & (4.48) \end{aligned}$$

$$\begin{aligned} &\int_R \log\left(1+\frac{\lambda}{\delta}x^2\right)\Pi_L(dx) < -\log\delta, & (4.49)\\ \Leftrightarrow\quad & EX_1 > 0, & (4.50)\\ \Leftrightarrow\quad & \sigma_t^2 \xrightarrow{d} \sigma_\infty^2 \overset{d}{=} \beta\int_0^\infty e^{-X_t}dt. & (4.51) \end{aligned}$$

Moving to other attributes of the stochastic-volatility process, it is suggested that mean can also vary with time (Balduzzi, Das and Foresi (1998)). In a context very similar to that one used for stochastic volatility models, they suggest modelling the underlying as below:

$$\begin{aligned} d\ln S_t &= \alpha\left(\mu_t - \ln S_t\right)dt + \sigma dW_{1t}, & (4.52)\\ d\mu_t &= \eta\left(\vartheta - \mu_t\right)dt + \gamma dW_{2t}. & (4.53) \end{aligned}$$

Brownian motions W_{1t}, W_{2t} are again correlated.

To all the above variations can very naturally follow. For example Bakshi, Cao & Chen propose extending Heston's model even more by incorporating jumps in the Stochastic Volatility process. Mencia and Sentana (2008) also propose a Central Tendency model (time-dependent mean) with jumps as

well as one with Central Tendency and Stochastic Volatility together.

Since analytical tractability is hard, Monte Carlo estimation has become a powerful and sometimes the only tool for every quest to overcome pricing obstacles. For that reason complicating the dynamics that drive the underlying asset (adding for example double mean reversion to the stochastic volatility of the underlying, Gatheral (2008) has become plausible but the extent to which such complications are necessary has yet to be answered.

What can be said, though, is that a catholic process that captures all modelling necessities dictated by contemporary market irregularities is still elusive. Distributions have become even more negatively skewed and take shapes that can make their identification even harder for the researcher (structural changes dependent on time). Thus the right to question the adopted distributional pattern is undisputable, while only one fact remains definitely true: conquering new grounds of scientific knowledge regarding the characteristics that drive the probability distribution of the underlying enables us to justify the pursuit of a better modelling proposal.

There is where model averaging can assist in that journey. The idea is simple and is based on the following. There might be the case that no single model can optimally capture all the attributes of the dependent asset process. Not only that, but also by combining all or just the top performing models together one can achieve even better forecasting performance (synergies are created). Researchers have proven empirically that forecasting accuracy when model averaging is applied can actually not only be improved but also we can accomplish success rates even higher than those single models can generate. Most popular methods used in the past to perform model averaging are Thick Modelling and Bayesian Model Averaging. While the first is straightforward, the second requires the unavoidable reference to Bayesian Theory, which can further place a computational burden to the estimation process. For that reason Bayesian Approximation is most commonly used instead, which also as a method avoids any inferences about prior and posterior distribution of the parameters. The posterior probability is simply approximated using information criteria. The steps followed to perform the averaging techniques discussed above will be presented analytically in later sections.

4.2.3. Transforming Risk-Neutral to Real-World Densities

As mentioned previously, Breeden and Litzenberger (1978) were the first to relate the option price to the risk-neutral density of the underlying. If we assume that the price is twice differentiable with respect to the strike, then the risk-neutral density $f^{Q}(x)$ can be derived using the following equation:

$$f_t^{Q}(S_T) = \frac{\partial^2 c(K)}{\partial K^2}, \tag{4.54}$$

where $c(K)$ is the option price and can also be expressed as the integral of the product of the option's payoff for a continuum of strike prices with its risk-neutral density.

$$\begin{aligned} c(K) &= E_t^{Q}\left[e^{-r(T)}(S_T - K)^{+}\right] \\ &= e^{-r(T)}\int_0^{\infty}(x-K)^{+} f^{Q}(x)\,dx \\ &= \int_0^{\infty}\zeta(x)(x-K)^{+} f^{P}(x)\,dx \\ &= E^{P}\left[\zeta(S_T)(S_T-K)^{+}\right], \end{aligned} \tag{4.55}$$

where $\zeta(x) = e^{-r(T)}\frac{f_t^{Q}(x)}{f_t^{P}(x)}$ is the stochastic discount factor.

Trying to interpret this approach, we notice that if we knew the option prices under all strikes for a particular date t, then we could derive the risk-neutral distribution of the underlying. This method, although it has been criticized due to its unrealistic assumption that option prices with a particular maturity exist for an infinite number of strikes still attracts the majority of interest due to its simplicity of implementation and its effectiveness in approximating the density of the underlying when working as closely as possible to the observed in the market strikes price range. Of course extending beyond that range is important.

Although the above proposition relates options to the risk-neutral and the real-world density of the underlying asset, it says nothing about investors risk preferences and how these could help in the transformation procedure from a $Q-$ to a $P-$measure. Aït-Sahalia and Lo (2000), followed by Jackwerth (2000) and Rosenberg and Engle (2002), made this connection:

$$f^P(S_T) = \frac{\frac{f^Q(S_T)}{U'(S_T)}}{\int_0^\infty \frac{f^Q(S_T)}{U'(x)}dx}. \tag{4.56}$$

Initially we have to assume a utility function $U(W_T)$, where W_t is the wealth at time T. We know that (Merton (1971)) the target of the investor is to maximize wealth. Assuming that we invest all our wealth in the risky asset such that at the end our wealth equals the asset price (i.e. $W_T = S_T$), then the stochastic discount factor for all options is a random variable that equals

$$\zeta(S_T) = e^{-r(T-t)}\frac{f^Q(S_T)}{f^P(S_T)} = \lambda\frac{U'(S_T)}{U'(S_t,)} \tag{4.57}$$

with $\frac{U'(S_T)}{U'(S_t)}$ the marginal rate of substitution (or else stochastic discount factor) and λ some constant the estimation of which can be disregarded from the analysis (as we will prove later) if particular utility functions are chosen to represent investor's risk appetite. The relative risk aversion is

$$RRA = -x\frac{\zeta_t'(x)}{\zeta_t(x)} = -x\frac{U''(x)}{U'(x)}, \tag{4.58}$$

and can be approximated as follows: [5]

$$\begin{aligned}
\widehat{RRA} &= -S_T\frac{\zeta_t'(S_T)}{\zeta_t(S_T)} \\
&= -S_T\frac{\frac{f^{Q'}(S_T)f^P(S_T) - f^Q(S_T)f^{P'}(S_T)}{(f^P)^2(S_T)}}{\frac{f^Q(S_T)}{f^P(S_T)}} \\
&= S_T\frac{\hat{f}_t^{P'}(S_T)}{\hat{f}_t^P(S_T)} - S_T\frac{\hat{f}_t^{Q'}(S_T)}{\hat{f}_t^Q(S_T)}
\end{aligned} \tag{4.59}$$

and

$$f^P(S_T) = \frac{\frac{f^Q(S_T)}{\zeta(S_T)}}{\int_0^\infty \frac{f^Q(x)}{\zeta(x)}dx}$$

[5] For more details refer to the original paper: Aït-Sahalia & Lo (2000)

$$= \frac{\frac{U'(S_t)}{\lambda U'(S_T)} f^Q(S_T)}{\int_0^\infty \frac{U'(S_t)}{\lambda U'(x)} f^Q(x)\,dx}$$
$$= \frac{\frac{f^Q(S_T)}{U'(S_T)}}{\int_0^\infty \frac{f^Q(S_T)}{U'(x)dx}}. \tag{4.60}$$

Let us now assume that the investor has the power utility function as a representative of his risk aversion. The power utility is defined as below:

$$U(x) = \begin{cases} \frac{x^{1-\gamma}}{1-\gamma} & \gamma \neq 1, \\ \log(x) & \gamma = 1. \end{cases} \tag{4.61}$$

Then the Relative Risk Aversion is

$$\widehat{RRA} = -x\frac{U''(x)}{U'(x)} = -x\frac{-\gamma x^{-\gamma-1}}{x^{-\gamma}} = \gamma \tag{4.62}$$

and

$$f^P(x) = \frac{x^\gamma f^Q(x)}{\int_0^\infty f^Q(y)\, y^\gamma dy}. \tag{4.63}$$

Finally, the integral in the above equation has to be evaluated numerically. However, the above is a pragmatic utility transformation method where the stochastic discount factor is proportional to the utility function.

According to Ritchey (1990), Melick and Thomas (1997) and Brigo and Mercurio (2002), the risk-neutral density of the asset price can be defined as a mixture of lognormal densities as well. In this case, it can be proven that the above transformation form the risk-neutral to the real-world density can be performed as follows:

$$f^P(x|\theta,\gamma) = f^Q(x|\tilde{\theta}) \tag{4.64}$$

with real-world parameters

$$\tilde{\theta} = (\tilde{F_1}, \tilde{F_2}, \sigma_1, sigma_2, \tilde{w}) \tag{4.65}$$
$$\tilde{F_i} = F_i exp(\gamma, \sigma_i^2 T), \quad \forall i = 1, 2, ... \tag{4.66}$$
$$\frac{1}{\tilde{w}} = 1 + \frac{1-w}{w}(\frac{F_2}{F_1})^\gamma exp(\frac{1}{2}(\gamma^2-\gamma)(\sigma_2^2-\sigma_1^2)T). \tag{4.67}$$

where F_i is the future price, and w the weight assigned to the first lognormal density (and equivalently $1-w$ is the weight assigned to the second lognormal distribution).

A similar expression can be written for the Generalizes Beta distribution of second type (GB2), first introduced by Bookstaber and McDonald (1987) and utilized by Anagnou et al. (2002). There are four positive parameters, $q = (a, b, p, q)$, involved in the specification of the distribution. When changing measures, from risk-neutral to real-wold

$$f^P(x|\theta,\gamma) = f^Q_{GB2}(x|\tilde{\theta}) \tag{4.68}$$

with real-world parameters

$$\tilde{\theta} = (a, b, p + \frac{\gamma}{a}, q - \frac{\gamma}{a}) \tag{4.69}$$

assuming $aq > \gamma$.

Except from the utility method, there is the statistical recalibration method, as this is presented in Liu, Shackleton, Taylor and Xu (2007), and in Shackleton, Taylor and Yu (2010).

4.3. Real-World Volatility Density Forecasting

4.3.1. Literature Review

Until now there has been an extensive reference regarding forecasting the risk-neutral density of the volatility of an asset as this is inferred from contingent claims written on its volatility index. However, there is an alternative choice of directly using the asset prices to make relevant inferences. To execute this method we only need to assume the process that drives the volatility of that asset (discrete-time model) and then use bootstrap techniques obtain the predictive volatility density according to different assumptions for the error distribution. Modelling the volatility of an index, for example, has been so extensively addressed in the past that there exists an abundance literature devoted to the subject. Without any detailed reference to all discrete-time conditional-volatility models, as this is provided in previous chapters, it suffices to say that the great predecessor of all is ARCH (Engle (1982)) and consequently GARCH (Bollerslev (1986)). Engle

in his model assumes that the variance of the error term is a function of the previous period error plus a constant, while Bollerslev adds to the equation the lagged value of that variance. More precisely, let y_t be a sequence of realizations such that

$$y_t = \mu_t(y_t) + \varepsilon_t, \tag{4.70}$$
$$\varepsilon_t = \sqrt{h_t} z_t, \tag{4.71}$$

where $z_t \overset{iid}{\sim} N(0,1)$. In turns the conditional variance h_t is defined as below.

Model	*Variance Equation*	*Restrictions to ensure positivity of variance*
ARCH	$h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2$	$\omega > 0$ $a_i \geq 0, i = 1, 2, \ldots p$
GARCH	$h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}$	$\omega > 0$ $a_i \geq 0,\ i = 1, 2, \ldots p$ $\beta_j \geq 0,\ j = 1, 2, \ldots p$ $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} \beta_j < 1$

Table 4.2.: The "work horses" of discrete-time conditional-volatility forecasting along with their restrictions.

The plethora of extensions that have followed stand as a testimony to the overwhelming interest that volatility forecasting attracts, as well as of the immense importance placed by researchers on having a model that tracks it correctly.

Indicatively, some of the most familiar models[6] include EGARCH (Nel-

[6] A good review is the following although the reader should refer whenever possible to original papers (see end references): Teräsvirta (2006)

son (1991)) which by modelling the log-variance removes any restrictions having to be imposed on the parameters, GJR GARCH (Glosten, Jagannathan, Runkle (1993)) that accounts for asymmetric impact on volatility of negative and positive shocks. A very close variation is where instead of variance the volatility is modeled asymmetrically, and this model is called TGARCH (Threshold GARCH, Zakoian (1994)), while a DTGARCH (Double Threshold, Li and Li (1996)) also exists. Another asymmetric variation comes from Engle and Ng (1993), in which the number of negative to positive stock responses is not equal. Other alternatives that account for leverage effects include Asymmetric Power GARCH (Ding, Granger and Engle (1993)) which in its more restrictive formation yields the Nonlinear GARCH (Higgins and Bera (1992)). In the literature we can also find from simple modelling solution such as the Absolute Value GARCH (Taylor (1986)) to more complicated structures like STGARCH (Smooth Transition GARCH, Hagerud (1997)) or EST (Exponential Smooth Transition, Taylor (2004)). Also long memory has been an indispensable part of volatility modelling, with references including Integrated and Fractional Integrated GARCH models (Nelson (1991)). Finally recent additions to the GARCH family models include the Time-Varying GARCH (Starica, 2004), where parameters in the variance equation depend on time, and also and the Markov Switching GARCH alternatives.

4.4. Density Forecasting Evaluation

Of the first to drew attention to the importance of density forecasting evaluation were Diebold and Lopez (1996), followed by Diebold Hahn and Tay (1998, 1999) and finally Diebold Gunther and Tay (1998). The great contribution of their research is a comprehensive measure for assesing the predictive density even when the underlying true distribution is unknown. This is of immense importance when the focal point of the researcher is volatility density forecasting, which by default implies that the true probability distribution is never observed. Also an additional advantage of the method is that it does not use any ranking criterion, which in turn could render the final comparative result biased towards the loss function used to rank the forecasts. According to the authors (Diebold, Gunther and Tay (1998)), let $\{y_t\}_{t=1}^{m}$ be a sequence of m realizations generated from a sequence of

conditional densities, $\{f_t(y_t|\Omega_t)\}_{t=1}^m$, where $\Omega_t = \{y_{t-1}, y_{t-2}, \ldots\}$ is the information set including all available past information and $\{p_t(y_t)\}_{t=1}^m$ are the relative density forecasts. Then the predictive distribution equals the true distribution of interest, $\{f_t(y_t|\Omega_t)\}_{t=1}^m = \{p_t(y_t)\}_{t=1}^m$, if and only if $\{z_t\}_{t=1}^m \overset{iid}{\sim} U(0,1)$, where z_t is the probability integral transformation of observed value with respect to the forecasted density $p_t(y_t)$ or else the forecasted cumulative density function $P_t(y_t)$ estimated at y_t

$$z_t = \int_{-\infty}^{y_t} p_t(u)du \equiv P_t(y_t). \tag{4.72}$$

Under the assumption of nonzero Jacobian matrix $\partial P_t^{-1}(z_t)/\partial z_t$ with continuous partial derivatives, z_t has the density function $q_t(z_t)$:

$$q_t(z_t) = \left|\frac{\partial P_t^{-1}(z_t)}{\partial z_t}\right| f_t\left(P_t^{-1}(z_t)|\Omega_t\right) = \frac{f_t\left(P_t^{-1}(z_t)|\Omega_t\right)}{p_t\left(P_t^{-1}(z_t)\right)}. \tag{4.73}$$

From the above equation it is obvious that if the predictive equals the real density then,

$$q_t(z_t) = \begin{cases} 1 & \text{if} \quad 0 \leq z_t \leq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{4.74}$$

where $z_t \sim U(0,1)$.

For the total sequence of forecasts, if $\{f(y_t|\Omega_t)\}_{t=1}^m = \{p_t(y_t)\}_{t=1}^m$, then the joint density is

$$q(z_m, \ldots, z_1) = \frac{f_m\left(P_m^{-1}(z_m)|\Omega_m\right)}{p_m\left(P_m^{-1}(z_m)\right)} \times \ldots \times \frac{f_1\left(P_1^{-1}(z_1)|\Omega_1\right)}{p_1\left(P_1^{-1}(z_1)\right)}, \tag{4.75}$$

where $q(z_m), \ldots, q(z_1)$ are the marginal uniform distributions such that $q(z_m, \ldots, z_1)$ is a multivariate uniform and $\{z_t\}_{t=1}^m \overset{iid}{\sim} U(0,1)$.[7]

In order to test for independence and uniformity a variety of alternatives exists in the literature. Diebold *et al.* refrained from using such tests, since

[7]For the proof of this equation refer to the original paper of Diebold, Gunther & Tay (1998).

if the initial hypothesis is rejected they do not provide in turn any further information regarding the reason of this outcome. Thus they are considered of limited value to the researcher. Instead they adopt a more simplistic graphical approach (histograms and correlograms). In other words, in order to reject or accept the null, some statistical tests tend to depend on examining different assumptions; thus the method they follow to reach a conclusion differs significantly between those indicators. But one cannot disregard the importance of these tests which have been used in the statistics domain and in empirical applications from academics and practitioners widely throughout the years. Among the most influential statistical tests for normality, uniformity and independence we can identify the following:

- Kolmogorov-Smirnov test (Massey (1951))

  This is a non-parametric distribution-free statistical test of the departure of the empirical CDF from the alleged distribution (usually it is the standard normal, but it could be used to test uniformity or any other continuous distributional pattern).

- Cramér-von Mises test (Cramér (1928), Von Mises (1931))

  This is an alternative to the Kolmogorov-Smirnov test of departure from a prespecified distributional assumption as opposed to the observed empirical distribution.

- Jarque-Bera test (Jarque and Bera (1987))

  This is related to assessing the normality assumption of the data by looking at the third and fourth moments of the sample. The t-statistic is assumed to be chi-square distributed.

- Diebold-Mariano test (Diebold and Mariano (1995))

  Let y_t denote the series to be forecasted and $\hat{y}^1_{t+h|t}$, $\hat{y}^2_{t+h|t}$ denote two competing forecasts of y_{t+h}. The forecast error for the two models is

$$\varepsilon^1_{t+h|t} = y_{t+h} - \hat{y}^1_{t+h|t} \tag{4.76}$$

$$\varepsilon^2_{t+h|t} = y_{t+h} - \hat{y}^2_{t+h|t}. \tag{4.77}$$

For the whole forecasting period $t = t_0, t_1, \ldots, T$ we have a different sequence of errors for the two approaches,

$$\{\varepsilon^1_{t+h|t}\}^T_{t_0},\ \{\varepsilon^2_{t+h|t}\}^T_{t_0} \tag{4.78}$$

In order to measure the accuracy of each forecast we use various loss functions for example,

- Squared error loss: $L(\varepsilon^i_{t+h|t}) = (\varepsilon^i_{t+h|t})^2$,
- Absolute error loss: $L(\varepsilon^i_{t+h|t}) = |\varepsilon^i_{t+h|t}|$,

and test the following null hypothesis:

$$\begin{aligned} H_0 :\ & E[L(\varepsilon^1_{t+h|t})] = E[L(\varepsilon^2_{t+h|t})], \\ H_1 :\ & E[L(\varepsilon^1_{t+h|t})] \neq E[L(\varepsilon^2_{t+h|t})]. \end{aligned} \tag{4.79}$$

Diebold-Mariano test in particular examines if the loss difference of the two different forecasting models has zero mean and unit variance,

$$d_t = L(\varepsilon^1_{t+h|t}) - L(\varepsilon^2_{t+h|t}), \tag{4.80}$$

while the null hypothesis can be expressed as follows:

$$\begin{aligned} H_0 :\ & E[d_t] = 0, \\ H_1 :\ & E[d_t] \neq 0. \end{aligned} \tag{4.81}$$

The test statistic is

$$t = \frac{\overline{d}}{(\widehat{avar}(\overline{d}))^{1/2}} = \frac{\overline{d}}{(L\hat{R}V_{\overline{d}}/T)^{1/2}}, \tag{4.82}$$

where

$$\overline{d} \ = \ \frac{1}{T_0}\sum_{t=t_0}^{T} d_t, \tag{4.83}$$

$$LRV_{\overline{d}} \quad = \quad \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j,\ \gamma_j = cov(d_t, d_{t-j}). \tag{4.84}$$

In order to avoid problems with serial correlation among the differentials, $\{d_t\}_{t_0}^{T}$, we have used above a more robust estimator for the variance $\sqrt{T}\ \overline{d}$, the $\widehat{LRV}_{\overline{d}}$ estimator. Under the null hypothesis the test statistic must be standard normally distributed (Diebold and Mariano (1995)),

$$t \sim N(0,1). \tag{4.85}$$

- Ljung-Box test

  This is a joint test that all autocorrelation coefficients equal zero up to a particular lag. Under the null hypothesis that the data are iid, the Q-statistic is chi-square distributed with ℓ degrees of freedom, $Q_{\substack{Ljung\\Box}} \sim \chi^2_{\ell}$.

- Berkowitz test (Berkowitz, 2001)

  The additional advantage of this indicator in comparison with the ones mentioned before is that both independence and uniformity are jointly tested here. The author starts with a further transformation of z_t (where z_t is defined as above by Diebold et al (1998)) by taking its inverse standard normal cumulative function value (y_t):

  $y_t = \Phi^{-1}(z_t)$.

  Under the null assumption y_t is distributed as follows,

  $y_t \overset{iid}{\sim} N(0,1)$.

  To test for normality and independence a likelihood-ratio test is used with critical values being taken from the chi-square distribution with two degrees of freedom, $LR_2 \sim \chi^2_2$.

4.5. Methodology

As discussed previously, we use two distinct approaches for volatility-density forecasting. One is based on deriving the risk-neutral density from option

prices written on a volatility index i.e. predicting the risk-neutral distribution of the underlying (volatility) and then transforming it to a real-world equivalent density. The second approach uses real-world data to forecast directly the volatility distribution by sampling multiple random draws for the error from a proposal distribution and then forecasting the variance. In more detail, the method used for both alternatives will be discussed in this part, which is naturally divided in two subsections depending on the adopted approach used for deriving the volatility density.

4.5.1. Real-World Volatility-Density Forecasting

Since models related to volatility forecasting are so numerous the researcher has to decide entirely within his discretion which is the most appropriate solution and with what specifications-lags. Model selection criteria do exist, but rarely provide a competing answer out-of-sample. These selection criteria make inferences based on in-sample information, which can only be slightly related to the uncertainty of any future outcome. Here we focus mainly on five out of all proposed models found in relevant literature, since incorporating everything in the analysis is not possible. In particular, the models that will be used along with their characteristic equations are provided below. It is worth noticing that these were selected either to be used as benchmarks or because they account for some of the most widely recognised characteristics of volatility.

Model	*Variance Equation*
GARCH	$h_t = \omega + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}$
Absolute Value GARCH	$\sqrt{h_t} = \omega + \sum_{i=1}^{p} a_i \left\|\varepsilon_{t-i} - c\right\| + \sum_{j=1}^{q} \beta_j \sqrt{h_{t-j}}$
GJR	$h_t = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \gamma \varepsilon_{t-i}^2 I\{\varepsilon_{t-i} < 0\} + \sum_{j=1}^{q} \beta_j h_{t-j}$
TGARCH	$\sqrt{h_t} = \omega + \sum_{i=1}^{p} a_i \left\|\varepsilon_{t-i} - c\right\| + \sum_{j=1}^{q} \beta_j \sqrt{h_{t-j}}$

Asymetric Power GARCH	$\left(\sqrt[\delta]{h_{t-j}}\right)^2 = \omega + \sum_{i=1}^{p} a_i \left\|\varepsilon_{t-i}\right\|^\delta - \gamma \left\|\varepsilon_{t-i}\right\|^\delta I\left\{\varepsilon_{t-i} < 0\right\} + \sum_{j=1}^{q} \beta_j \left(\sqrt{h_{t-j}}\right)^\delta$

Table 4.3.: Table of the GARCH-type model alternatives used for the estimation of the real-world volatility-forecasting density.

Also the error parameters are allowed to be distributed less restrictively including along with the Normal Distribution the Generalized Normal Distribution, the Student-t distribution and finally the Skew-t distribution. In more detail some basic characteristics of the functions are provided below. This is all in an attempt to better approximate the error distribution, since it is proven empirically that returns are leptokurtic with non-constant variance dependence.[8]

[8]See: Hansen (1994), Patton (2010), Sheppard (2010)

Distribution	*Expression*	*Parameter Bounds*	*PDF*
Normal	$z_t \sim N(\mu, \sigma)$	μ : mean $\sigma \geq 0$: volatility	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(z_t-\mu)}{2\sigma^2}\right)}$
GED	$z_t \sim GED(\mu, \alpha, \beta)$	μ : mean α : scale parameter β : shape parameter	$\frac{\beta}{2\alpha\Gamma\left(\frac{1}{\beta}\right)} e^{-\left(\frac{\lvert z_t-\mu\rvert}{a}\right)^{\beta}}$
Student-t	$z_t \sim ST(\nu)$	$2 < \nu < \infty$ degrees of freedom	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{z^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$ where Γ is a Euler's Gamma function.
Skew-t	$z_t \sim SKST(\nu, \lambda)$	$2 < \nu < \infty$ degrees of freedom $-1 < \lambda < 1$ skewness parameter	$\begin{cases} bc\left(1 + \frac{1}{\nu-2}\left(\frac{bz_t+\alpha}{1-\lambda}\right)^2\right)^{-\frac{(\nu+1)}{2}} & z_t < -\frac{\alpha}{b} \\ bc\left(1 + \frac{1}{\nu-2}\left(\frac{bz_t+\alpha}{1+\lambda}\right)^2\right)^{-\frac{(\nu+1)}{2}} & z_t \geq -\frac{\alpha}{b} \end{cases}$ where: $\alpha = 4\lambda c\left(\frac{\nu-2}{\nu-1}\right)$ $b = \sqrt{1 + 3\lambda^2 - \alpha^2}$ $c = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma\left(\frac{\nu}{2}\right)}$

Table 4.4.: Table of the error distributions used in the analysis.

Using real-world data to forecast the distribution of returns has been discussed by Rosenberg and Engle (2002). The method they used can be described as follows. Firstly they sample random independent innovations, and secondly for every draw they re estimate the return such that at each particular time point there exists a collection of simulated return estimates. What they suggest, however, is sampling from the empirical error distribution draws and not assuming any particular underlying density pattern (similarly with bootstrap methods).[9] Taylor (2002) also provides an example of forecasting the density of an equity index, but instead uses standardized random numbers which are drawn from a suggested predictive distribution $\mathrm{F}D\left(0,1\right)$. After all random draws are acquired they are used to estimate the conditional mean and variance in order to forecast the desired density.

Moving forward, we focus here on volatility density forecasting, where at each time step we first have to calibrate the parameters of the variance equation based on past information, $h_t\left|\Omega_{t-1}\right.$, where Ω_{t-1} is the information set at time $t-1$. For example, let us assume that we want to forecast the volatility distribution at time t. We achieve that by generating $h_{t,i}\ \forall i=1,2\ldots,n$ independent volatility forecasts with marginal distribution the same as that selected for the error. More precisely, at our discretion we have all returns from time 1 to time $t-p\ \forall p=1,\ldots,4$, where p is lag (here we consider a maximum lag of four, so p will take values from 1 to 4). At first we have to select one particular model to capture volatility along with the number of lags we want for that model (p,q). Secondly using maximum likelihood and assuming that the noise term in the variance equation takes one of the four proposal distributions (Normal, GED, Student t and Skew t) we are able to calibrate the parameters based on past information. Finally, after parameter estimates have been acquired we are ready to perform volatility-density forecasting. All that is required is to simulate a sufficiently large number of variance forecasts for time t by taking n independent random draws from the error distribution and adding them to the volatility equation. In other words, for each draw $z_{t-p,i}$ where $i=1,2\ldots,n$ we estimate $\varepsilon_{t-p,i}\ \forall i=1,2\ldots,n$ and next forecast $h_{t,i}\ \forall i=1,2\ldots,n$ (where n is the number of simulations). At the end, we have a volatility density estimate

[9] A good reference for the bootstrap methods is: Efron & Tibshirani (1993).

for time t for every univariate conditional volatility model with lags (p, q) and marginal distribution the same one as that of the error distribution. In steps the procedure is described as follows.

1. Select a model to forecast volatility and select an error distribution to use for that model.

2. Using return data until $t - p$ (p being the lag value) we calibrate volatility parameters using maximum likelihood. Notice that depending on the error distribution the log-likelihood considered necessary to calibrate that volatility model should be adjusted accordingly.

3. Take n random draws from that selected error density to simulate values for z_{t-p}.

4. Estimate $\varepsilon_{t-p,i} = \sqrt{h_{t-p,i}} z_{t-p,i} \quad \forall i = 1, 2 \ldots, n$

5. Create n independent volatility forecasts for each time step using the simulated error values.

6. Fit the volatility sample to a choice of three distributions. First to the empirical distribution using a kernel regression, second to the Gamma and third to the Log-Normal distribution.

7. Assess the accuracy of each model using the Berkowitz test for the whole sequence of forecasts. The model that generates the highest p-value for that test is the optimal model to forecast the volatility density of the index.

Regarding the reliance on p-values for determining the optimal density choice we have to notice the following. This is considered an appropriate path to follow when we can clearly accept one competing alternative and reject all the rest. However, when this acceptance/rejection is not catholic, and multiple models reject/accept the null hypothesis for a particular confidence level, the answer to which is the most dominant is not straightforward. Here, following Bliss and Panigirtzoglou (2002), we select among these densities that surpass the statistical threshold, the one that has the most satisfactory p-value, or else this distribution alternative that most "strongly" rejects the null hypothesis. Further consideration of this approach however, is of course suggested.

4.5.2. Risk-neutral Volatility-Density Forecasting (Option-based approach)

In order to acquire the predictive risk-neutral volatility density we have to acquire relative option prices for a continuum of strikes, all with the same time to maturity. The models that are used to facilitate the estimations along with some intrinsic details of their implementation are provided below.

- The analytic pricing formula of Grünbichler and Longstaff, a Square-Root Mean-Reverting Process where according to their assumptions volatility is noncentrally $\chi^2(\nu, \lambda)$ distributed. In particular the discretized analytical expression of the pricing formula is given below.

$$\begin{aligned} c(V_t, K) &= e^{-rt}[e^{-\beta t} V_t Q(\gamma K|\nu+4, \lambda) + \frac{\alpha}{\beta}(1-e^{-\beta T})\ldots \\ &\quad Q(\gamma K|\nu+2, \lambda) - e^{-rT} K Q(\gamma K|\nu, \lambda)], \end{aligned} \tag{4.86}$$

where as defined before,

$$\nu = \frac{4\alpha}{\sigma^2}, \tag{4.87}$$

$$\lambda = \gamma e^{-\beta T} V, \tag{4.88}$$

$$\gamma = \frac{4\beta}{\sigma^2 (1-e^{-\beta T})}. \tag{4.89}$$

$$\tag{4.90}$$

$Q(\cdot|\nu, \lambda)$, whose definition has been provided above as the complementary distribution function of the noncentral $\chi^2(\nu, \lambda)$, can be approximated as follows.

$$Q(\gamma K|\nu, \lambda) = 1 - N(d, l, \kappa), \tag{4.91}$$

where $N(d, l, \kappa)$ is the Normal CDF function with mean l and variance κ and can be estimated as below:

$$h = 1 - \frac{2}{3}(\nu+\lambda)(\nu+3\lambda)(\nu+2\lambda)^{-2}, \tag{4.92}$$

$$\kappa = \left[h^2 \frac{2(\nu+2\lambda)}{(\nu+\lambda)^2}\left[1-(1-h)(1-3h)(\nu+2\lambda)(\nu+\lambda)^{-2}\right]\right]^{-\frac{1}{2}}, \tag{4.93}$$

$$l = 1 + h(h-1)\frac{\nu+2\lambda}{(\nu+\lambda)^2} - h(h-1)(2-h)(1-3h)\frac{(\nu+2\lambda)^2}{2(\nu+\lambda)^4}, \tag{4.94}$$

$$d = \left(\frac{\gamma K}{\nu+\lambda}\right)^h. \tag{4.95}$$

- The Log-Mean-Reverting Gaussian Process of Detemple and Osakwe, which is also analytically tractable with solution

$$V_T = e^{-rt}V_0^{\phi_t} e^{\frac{\alpha(1-\phi_t)}{\lambda}+\frac{\eta^2}{2}} N(\delta+\eta) - KN(\delta), \tag{4.96}$$

where $N(\cdot)$ is the standard normal distribution, while the rest are defined below.

$$\phi_t = e^{-\lambda t}, \tag{4.97}$$

$$\eta = \frac{\sigma\sqrt{(1-\phi_t^2)}}{\sqrt{2\lambda}}, \tag{4.98}$$

$$\delta = \frac{1}{\eta}\left(\phi_t \log(V_t) - \log(K) + \frac{\alpha(1-\phi_t)}{\lambda}\right). \tag{4.99}$$

- The risk-adjusted Jump-Diffusion Mean-Reverting model of Clewlow and Strickland which is discretized as below.

$$\Delta \log V_t = \left(\alpha(\mu - r - \log V_t) - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}\varepsilon_{1t} + \ldots \\ (\mu_{jump} + \sigma_{jump}\varepsilon_{2t} - \kappa\mu_{jump}\Delta t)(u_t > \kappa\Delta t), \tag{4.100}$$

where r is the risk premium, μ_{jump} and σ_{jump} is the mean and the variance of the jump process and κ is the jump-frequency parameter.

Additionally $\varepsilon_{1t},\ \varepsilon_{2t}$ are standard normal random numbers while u_t represents random draws from the uniform distribution.

$$\begin{aligned} \varepsilon_{1i}, \varepsilon_{2i} &\sim N(0,1) \\ u_i &\sim U(0,1) \text{ where } u_i \begin{cases} 1 & u_i > \kappa \Delta t, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{4.101}$$

- By intergrading Schwartz's (1997) model (non-constant volatility component for the equation that describes the underlying) with Detemple's and Osakwe's (2000) (log-Ornstein-Uhlenbeck model), we employ a Stochastic-Volatility double-Mean-Reversion process (both the underlying and the volatility process mean-revert). Formulating this in order to simulate the processes we have the following relationships.

$$d \log V_t = \alpha (\mu - r - \log V_t)\, dt + \sigma_t dW_{1t}, \tag{4.102}$$

$$d \log \sigma_t^2 = \eta \left(\vartheta - \log \sigma_t^2\right) dt + \gamma dW_{2t}. \tag{4.103}$$

Discretizing the above SDEs, we get

$$\Delta \log V_t = \left(\alpha (\mu - r - \log V_t) - \frac{1}{2}\sigma_t^2\right) \Delta t + \sigma_t \sqrt{\Delta t}\varepsilon_{1t}, \tag{4.104}$$

$$\Delta \log \sigma_t = \left(\eta (\theta - \log \sigma_t) - \frac{1}{2}\gamma\right) \Delta t + \gamma \sqrt{\Delta t} \left(\rho \varepsilon_{1t} + \sqrt{1-\rho^2}\varepsilon_{2t}\right). \tag{4.105}$$

- Another alternative in relation to the Barndorff -Nielsen-Shephard model (2001) is to capture the driving dynamics of volatility options using a Gamma Ornstein-Uhlenbeck process with stochastic time change. By making time stochastic we try to capture the fact that in periods of high volatility, time will run faster than in periods of low volatility. This potentially can impact the total return measured over one calendar day which can equal several days' return counted in the stochastic business time. In order to simulate such a process, we first

have to draw our attention to the volatility equation. This requires to define $y_t,\ t \geq 0$, as the rate of time change described by the following SDE:

$$dy_t = -\lambda y_t dt + dz_{\lambda t}, \tag{4.106}$$

where z is a compound Poisson process. Since time needs to increase all processes modelling the rate of time change need to be positive (Clark, 1973). Equivalently we have

$$y_{t_n} = (1 - \lambda\Delta t)\, y_{t_n - 1} + \sum_{\kappa = N_{t_{n-1}+1}}^{N_{t_n}} x_\kappa e^{-\lambda \Delta t u_\kappa}, \tag{4.107}$$

where u_κ are independent random draws from a uniform distribution, $u_\kappa \overset{iid}{\sim} U(0,1)$, $t_n = n\Delta t$ and $x_\kappa = -\ln(u_\kappa)/b$

Afterwards we are ready to return to the pricing equation which is discretized as before:

$$\Delta \log V_t = \left(\alpha\left(\mu - r - \log V_t\right) - \frac{1}{2}\sigma_t^2\right)\Delta t + \sigma_t \sqrt{\Delta t}\varepsilon_{1t}. \tag{4.108}$$

- Central Tendency (time dependent mean) with Jumps

  Discretizing in relation to what has been mentioned before, we get the following equations:

$$\mu_t = c + b\mu_{t-1} + \delta\varepsilon_t, \tag{4.109}$$

where

$$c = \theta\left(1 - e^{-\alpha\Delta t}\right), \tag{4.110}$$

$$b = e^{-\alpha\Delta t}. \tag{4.111}$$

The solution of the above system produces the following estimates for the parameters:

$$\alpha = -\frac{\ln(b)}{\Delta t}, \tag{4.112}$$

$$\theta = \frac{c}{1-b}, \tag{4.113}$$

$$\sigma = \frac{\delta}{\sqrt{(b^2-1)\,\Delta t/2\ln(b)}}. \tag{4.114}$$

In turn, the underlying follows an OU jump process with time-dependent mean, which is simulated using the equation

$$\begin{aligned}\Delta \log V_t &= \left(\alpha\left(\mu_t - r - \log V_t\right) - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}\varepsilon_{1t} \\ &+ \left(\mu_{jump} + \sigma_{jump}\varepsilon_{2t} - \kappa\mu_{jump}\Delta t\right)(u_t > \kappa\Delta t)\end{aligned} \tag{4.115}$$

All the above models can be used to approximate the risk-neutral density of the underlying as indicated by Breeden and Litzenberger. The target, though, is the true probability distribution, which as opposed to the GARCH methodology cannot be inferred directly from the observed option prices. For that reason the acquisition of the risk coefficient, necessary for the relevant transformation from the risk-neutral to the objective measure, is also an issue to be considered. We follow the Bliss and Panigirtzoglou method where the risk premium is obtained by maximizing the Berkowitz test p-value related to the realizations of the index with respect to the transformed predictive distribution. In other words, a series of real index prices is transformed into probabilities, using all the competing forecasting densities. A Berkowitz test is performed afterwards, while the risk premium is calibrated by optimizing the p value of the test.

In steps to acquire the predictive density we execute a series of calculations all presented below.

1. Select an option pricing model. To avoid problems related to model uncertainty later we will use model-averaging procedures instead of making such a decision.

2. Calibrate the parameters of the model using observed option prices.

3. Using a butterfly spread to approximate the 2nd derivative of the

price with respect to strike, obtain the empirical risk-neutral predictive distribution of the underlying (the VIX index) for a month ahead.

4. Obtain random draws from the empirical predictive density in order to calibrate the Gamma and the Log-Normal distribution that we want to use in addition to the empirical.

5. Transform the risk-neutral distribution (either the empirical, Gamma or Log-Normal) to the objective one according to the Bliss and Panigirtzoglou method and connect the investor's risk preferences (using the power utility function for example) with the risk-neutral real-world density.

6. Repeat steps 1 to 4 for the whole sequence of forecasts.

4.6. Density model averaging

Although averaging is not a novel idea the utilization of the method's advantages has to the best of our knowledge never been extended to density forecasting. Here we try to account for that gap in the literature by taking some of the most popular methods regarding averaging and use them when performing density forecasting.

These methods relate mainly to Bayes theory, while the technique used to account for the inherent uncertainty when model selection is performed is referred to as Bayesian Model Averaging (BMA). Here in particular we will employ an approximation of BMA, which is widely used by practitioners in order to avoid inherent problems related to the application of the method as already discussed in previous chapters. These problems are mainly the inferences required for the prior and the posterior distribution of the parameters, as well as further technicalities regarding the implementation of the theory. Bayesian Approximation (BA) uses the Schwartz Bayesian Information Criterion (SIC or BIC respectively) to approximate the Bayes Factors needed in order to perform averaging. Under this assumption no further reference is to the prior or posterior densities required and the researcher has a fast and reliable formula to use in the averaging procedure. Here we also extend this proposition by using not only the BIC but also the Akaike Information Criterion (AIC) as well as the Mean Squared Error (MSE) (or

any other ranking measure could be used as well). Further details regarding how Bayesian Approximation is performed when dealing with density forecasting will be provided below. Depending on the targeted distribution however (real-world or risk-neutral) the method is adapted accordingly.

Another popular technique that will be used for averaging the predictive densities is Thick Modelling or Thick Model Averaging (TM or TMA by Granger and Jeon, 2004). Thick Modelling also uses information criteria or any other ranking measure, as will be exploited here to rank competing models. Next the method takes only a percentage of the most successful competing alternatives based on their performance according to that error measure and places equal weights on those to generate the final averaged scheme. The method can be criticized as being dependent on the ranking criterion as well as on the selection of the percentage of top-performing models which dictates which models will be used to perform forecasting while the rest are discarded. It is empirically justified, however, that these are addressable uncertainties (as we have seen in the previous chapter as well as from what has been justified by Zaffaroni and Pesaran, 2006), while the dynamic of the method is undisputable. In most cases information criteria tend to give relevant rankings, while the same holds when using any error-ranking measure. Also empirical evidence supports the fact that smaller retention percentages between 10%-30% in the majority of cases generate optimal results. This percentage depends also on the number of model alternatives considered for averaging since the fewer the methods the higher the percentage. As a rule of thumb, however, the predetermined range of 10% to 30% is the most favourable percentage in practice.

4.6.1. Risk-Neutral Volatility-Density Averaging

When the risk-neutral volatility density is investigated the averaging relates to the option prices. In other words, instead of assuming one single model that we have to select in order to price options, we opt for an averaged scheme instead. Here a collection of competing pricing proposals is used in an attempt to address the problem of model uncertainty inherited in all financial applications. As discussed above, in this analysis we are considering six different models to price options, two which have analytical solutions (the proposal of Longstaff and Grünbichler first and that of Detemple and

Osakwe second) and an additional four alternatives which have to be approximated numerically. For each period these models are fitted to the real option data and the Mean Squared Error of that fit is reported. As is obvious, this MSE is related to the in-sample performance of those pricing models which are used as a starting point in order to apply the averaging techniques.

The formulae used to estimate the Akaike and Bayesian Information Criterion related to averaging methods for the risk-neutral volatility density forecasting are those presented below:

$$AIC_t = e_t^{\frac{2k}{N}} \frac{1}{N} \sum_{n=1}^{N} e_{n,t}^2 = e_t^{\frac{2k}{N}} MSE_t, \tag{4.116}$$

$$BIC_t = N^{\frac{k}{N}} \frac{1}{N} \sum_{n=1}^{N} e_{n,t}^2 = N^{\frac{k}{N}} MSE_t, \tag{4.117}$$

where e^2 is the residual vector or else the option pricing error generated from the fitted option prices to the real observations, N the sample size of the surface and k the number of parameters each model has. Also we have defined MSE as

$$MSE_t = \frac{\sum_{n=1}^{N} \left(C_{n,t} - \hat{C}_{n,t} \right)^2}{N} = \frac{\sum_{n=1}^{N} e_{n,t}^2}{N} \tag{4.118}$$

From the above equations the relationship between the two model selection measures and the Mean Squared Error is also apparent. However, these criteria account not only for the goodness of fit as the MSE itself can reveal but also for the increase of the estimation error caused by the addition of an extra parameter. A heftier penalty is intuitively applied when using the information criteria, and especially BIC, which produces more parsimonious models and for that reason it is favoured by researchers.

To apply Bayesian Approximation, all models will be incorporated in the averaged scheme, weighted according to the Bayes Factor related to every model. The Bayes factor according to Kass and Wasserman (1995) can be closely approximated by the Bayesian Information Criterion. The error of this approximation is sufficiently small for large samples. As a consequence of what has been mentioned above, the relevant weights that

have to be assigned to each option model in order to perform averaging are calculated using the equations provided below. Each one of them depends on the error criterion the researcher uses to rank the performance of the competing pricing schemes in-sample (AIC, BIC or MSE).

$$BF_{M_{i,t}} \approx \frac{BIC_{M_{i,t}}}{\sum_{i=1}^{n} BIC_{M_{i,t}}} \quad \forall i = 1, \ldots, n, \tag{4.119}$$

$$BF_{M_{i,t}} \approx \frac{AIC_{M_{i,t}}}{\sum_{i=1}^{n} AIC_{M_{i,t}}} \quad \forall i = 1, \ldots, n, \tag{4.120}$$

$$BF_{M_{i,t}} \approx \frac{MSE_{M_{i,t}}}{\sum_{i=1}^{n} MSE_{M_{i,t}}} \quad \forall i = 1, \ldots, n, \tag{4.121}$$

where n here is the number of available option pricing models and M_i is the model under consideration each time.

Thick Model averaging, on the other hand, uses only a percentage of the competing models, those that were more successful based again on those three ranking goodness of fit measures. For the risk-neutral volatility density the number of option-pricing alternatives is small, so the percentage remains at this instance only is chosen to equal to 33% and 50%. Normally when the total of competing models is large, smaller percentages should be chosen instead.

$$w_{M_{i,t}} = \frac{\sum_{i=1}^{p} BIC_{M_{i,t}}}{p} \quad \forall i = 1, \ldots, p, \tag{4.122}$$

$$w_{M_{i,t}} = \frac{\sum_{i=1}^{p} AIC_{M_{i,t}}}{p} \quad \forall i = 1, \ldots, p, \tag{4.123}$$

$$w_{M_{i,t}} = \frac{\sum_{i=1}^{n} MSE_{M_{i,t}}}{p} \quad \forall i = 1, \ldots, p, \tag{4.124}$$

where p is the number of models included in the averaging scheme depending on the percentage retained (so here $p = 33\%n$ or $p = 50\%n$ with n being, as defined above, the total of all models).

4.6.2. Real-World Volatility-Density Averaging

When considering forecasting the volatility distribution directly from the market realizations the averaging is performed to the forecasted distributions themselves and not to the option prices, as has been suggested in the previous section. The same averaging techniques will be used to make inferences about the predictive density here, also including BA and TMA. For that reason the steps followed include first the estimation of the in-sample goodness of fit measures i.e. the AIC, BIC and MSE. To calculate the above the output of the GARCH type models which were estimated using the maximum likelihood method is used. More precisely the information criteria are estimated this time using the following equations:

$$AIC_{M_t} = -2L_{M_{i,t}} + \frac{2}{N}k_{M_{i,t}} \quad \forall i = 1, \ldots, n, \tag{4.125}$$

$$BIC_{M_t} = -2L_{M_{i,t}} + \frac{\log(N)}{N}k_{M_{i,t}} \quad \forall i = 1, \ldots, n, \tag{4.126}$$

where n the number of predictive density models under consideration while k is the total number of parameters each volatility model has. In turn the MSE now equals:

$$MSE_{M_t} = \frac{\sum_{i=1}^{N}\left(h_{M_{i,t}} - \hat{h}_{M_{i,t}}\right)^2}{N} = \frac{\sum_{i=1}^{N} e_{M_{i,t}}^2}{N} \quad \forall i = 1, \ldots, n, \tag{4.127}$$

where N is the number of observations that are fitted in-sample, h the actual and $\hat{h}$ the forecasted variance derived from the M_i model. Finally the steps followed to apply Bayesian Approximation and Thick Model Averaging are the same as those described in the previous section.

4.7. Data

The data used for the empirical analysis are divided into two subgroups, depending on the targeted distribution and the method used to derive that. For the risk-neutral volatility-density forecasting, option surfaces related to the VIX index from 03/2006 until 10/2009 have been downloaded. As in the previous chapter, the source for the option data is Optionmetrics and it

was accessed from WRDS, while to acquire realizations of the S&P500 and VIX index the CRSP Database was used. This yields a total of 44 surfaces while the frequency used is monthly. This implies that the first day of every month a new option surface is acquired, and based on that a forecast for the next month's risk-neutral volatility density is made. However, extra care has been taken regarding the fact that the risk premium has to be recalibrated on a monthly basis (following Bliss and Panigirtzoglou), and also adequate historical information is required every time for the estimation process. For that reason the first 32 surfaces are discharged from the sample and are merely used to calibrate the risk premium. Only the last 12 surfaces are used at the end for density forecasting and evaluation of performance. Additionally a moving window for the estimation of risk premium is used, meaning that every month a new surface of option prices is added to the sample while the first surface from the start is discharged. At the end the size of the window remains constant to 32 months but the data pool is retained updated. After we recalibrate (on the first day of every month) the coefficient of risk aversion, we use the option prices of this particular month we are at to forecast the risk-neutral volatility density for next month.

Finally it has to be highlighted that not all option prices available every month are used. Based on a filtering process relevant to the one applied in the first chapter, we discard options with small maturity (less than ten days), as well as those with more than 365 days maturity. Additionally, options with value smaller than \$0.05 are sorted out and finally those that have zero volume. Table 4.17 summarizes how many options are left to work with for each surface after the filtration process.

For the second approach where GARCH-type models are used, inferences for the predictive volatility density are made directly from the index itself. The S&P 500 index is used while frequency remained monthly as in the previous approach. To estimate conditional volatility it is widely accepted that the more data available the better for GARCH type models, so the sample period starts from as far as 3/1950 and ends in 10/2009. This generates 716 monthly data points, which in turn are divided into the estimation period and out-of-sample period. The first consists of 704 observations while the remaining 12 realizations constitute the out-of-sample period. To evaluate the forecasting accuracy of every modelling alternative the high-low

volatility proxy is used.[10]

4.8. Results

4.8.1. Results of Risk-Neutral Volatility-Density Forecasting

Results are based on the performance of the models for the total sequence of forecasts (12, of monthly frequency) as this is determined using the Berkowitz test. In more detail, a month after the predictive density has been generated the realization of the index is reported. This realization is afterwards transformed into a probability, using the forecasted density produced a month before. This procedure is repeated for the total of 12 steps ahead forecasting horizon, and all those probabilities are reported. Finally, this probability sequence is used to perform the Berkowitz test. The optimal model is the one that produced the highest test p-value, or else the highest acceptance rate that the forecast distribution is the actual data generating density.

Regarding risk-neutral volatility density forecasting, the empirical results (Table 4.6)[11] indicate a catholic dominance of the averaging schemes. The majority of single model alternatives exhibit poor predictive power, while the p-values related to the Berkowitz test for these models are significantly smaller than those generated for the averaging alternatives.

However, model selection criteria, which serve as a tool which facilitates the decision-making process, are calculated during the estimation period. Empirical evidence suggests that this process fails to identify a modelling scheme that will be of the same competence out-of-sample. The researcher is left to make a conscientious choice regarding which alternative will outperform the other candidate proposals, substantiating his argument with the use of a goodness-of-fit measure. As we have seen, this widespread practice involves using information criteria or error measure criteria to assess the in-sample robustness of the competing models and to rank them accordingly. However, this is not the case in our empirical analysis. Results suggest that by using either the AIC, BIC or MSE to select the best alternative to make density forecasts we have ultimately a very poor out-of-sample predictive

[10] See third chapter section 5 for further details of the high-low volatility proxy

[11] For the total ranking tables that include all available single models and averaging processes see Table 4.10.

outcome. In other words, regardless of the ranking indicator, in-sample comparison of competing models unavoidably fails to provide a coherent answer to the forecasting problem and so leads to over-fitted incompetent selections.

More precisely, as far as the forecastability of the risk-neutral density is under scrutiny, the top performer is Thick Model Averaging. This choice can be made irrespective of the ranking criterion or the percentage retained (needed to be selected in order to apply the method), while the assumption of either a Log-Normal or an empirical distributional pattern for the volatility process is a preferable option (as opposed to a gamma equivalent).

Bayesian approximation also is useful when trying to predict the future risk-neutral volatility density, while similar performance is demonstrated when using the method under all three ranking criteria (AIC, BIC, MSE). However, the Log-Normal assumption of the volatility is proven empirically to work better under this averaging method.

Regarding the performance of the single models we notice the following. The highest p-values are recorded for option pricing methods that assume a process for the underlying that exhibits either jumps (with constant mean and variance) or has a stochastic volatility dynamic (without jumps). What seems the least preferable modelling alternative is the BNS model with a gamma stochastic-volatility mean-reverting process. Finally, on average the performance of single models when these are selected in-sample using the MSE ranking measure exhibit out of sample a better forecasting power than their equivalents under information criteria for model selection purposes. Additionally for the single models the Log-Normal or the empirical density are still better assumptions than the gamma distribution for the volatility. All in all, however, as mentioned before, single models fail to solidify their status as trustworthy density forecasting vehicles.

Moving to the transformed risk-neutral to objective volatility density (Table 4.7)[12] we notice the following. First and foremost, we observe a considerable drop of the p-value of the Berkowitz statistic for all models. Still it is apparent that the averaging processes prevail over single models. In general Thick Modelling under all different ranking criteria and retention percentages is ranked first when volatility is assumed to be distributed either Log-Normally or when we use its empirical density as well as under

[12]Table 4.11 includes the rankings of all models estimated.

the gamma distributional assumption. In other words, TM is a consistent performer under all different specifications.

Regarding the performance of single models, however, we observe a shift in the preferences when the risk premium is incorporated into the calculations. As a result when the Gamma distribution is used as an assumption for the volatility density, single processes exhibit a stronger acceptance of the Berkowitz test. In addition, using BIC to identify the most dominant model in-sample that will be used to make inferences for the future proves to be a stronger criterion than the other two model selection measures (AIC and MSE). On the contrary, under the risk-neutral measure it is the MSE using the Log-Normal or the empirical distributional assumption that is preferable, so inconsistences are reported here.

In addition, for single models we notice now the analytical option-pricing formula of Grünbichler and Longstaff, when used to derive the predictive density of the underlying, generates highest acceptance rates of the Berkowitz test. Worst choice is the model with jump diffusion and stochastic volatility assumption for the underlying.

In conclusions we can firmly assert that results indicated that Thick Model Averaging exhibits not only a stronger acceptance of the Berkowitz test but most importantly a consistent outperformance of single models when forecasting both the risk-neutral and the transformed objective volatility density. For that reason it should be chosen by researchers trying to account for the problem of model uncertainty, as this is validated by empirical evidence.

4.8.2. Results: Real-World Volatility-Density Forecasting

First, we have to notice that the realized real-world volatility is approximated using the high-low estimator (the same proxy has been used in the previous chapter (equation 3.43), since as we have seen volatility in general is an unobserved process. When the volatility density of the S&P 500 index is under investigation, the empirical analysis demonstrated the following (Table 4.5).[13]

The overwhelming majority of the methods fail to pass the Berkowitz test and produce statistically significant results when we assume that the volatility is Gamma distributed. On the contrary, the use of a Log-Normal

[13]Tables 4.12, 4.13 and 4.14 include the collective rankings of all models estimated.

volatility density instead produces higher acceptance rates of the null for most of the competing modelling alternatives. The only exception is reported at the top ranking positions when using the empirical distribution and the APGARCH model to derive the conditional volatility density. [14] Also notable is the fact that selecting the best model in-sample to make forecasts based on a goodness-of-fit measure (here we used information criteria) fails to pass the Berkowitz test and reports the lowest p-value in comparison with all other proposals (some though are equally worse none of those worst performers is an averaging scheme).

Thick Model Averaging, on the other hand, with higher retention rates (20%, 30%, 40%, 50%) and with the use of any of the AIC, BIC for ranking purposes has the highest acceptance of the null hypothesis (that the predictive density is the actual volatility generating distribution). Averaging here eliminates the indecisiveness regarding which error distribution should be adopted, as well as which single conditional volatility model can be used to capture more adequately the attributes of the index's variance distribution. Finally, BA in this part, although it produces better performance when compared to the rest of single models, does not necessarily generate significant results.

4.8.3. Investment Strategies

The investment strategy that is adopted in this chapter relates only to the option-based predictive-volatility density. Since VIX is considered by a lot of practitioners as an indicator of future movements of the S&P500, we used the forecasted distribution of VIX to predict the equity index. There are no short sales restrictions under our regime, while the transaction costs are considered negligible. The strategy is as follows.

First we take the current price of VIX and using the different predictive densities we obtain the cumulative probability. This probability can be interpreted as the confidence level we have that the same VIX price will appear next month. If the probability surpass a certain level it is an indication that with an accepted certainty VIX next month will be at the same or lowerhigher level, so we take a short position on the S&P500 index. On the contrary if it fails to exceed that threshold we take a long position on

[14] For the total ranking tables that include all available single models and averaging processes see in the appendix Tables 5.26, 5.27,5.28.

the index. The certain benchmark level is obtained dynamically by optimizing the cumulative wealth of the strategy described above, over the last 5 months. This is a dynamic approach where at each step the threshold should be recalibrated, so offering the researcher an updated information environment where decisions can be made with more assertiveness.

Results (Tables 4.8, 4.9) [15] indicate a higher cumulative wealth under the Bayesian Approximation scheme against all other alternatives. Thick Modelling on the contrary exhibits equal or superior performance when compared to some single model suggestions. All in all, TM exhibits on average satisfactory performance. In conclusions all averaging schemes are preferable to investing directly on the S&P500 index, while the worst cumulative wealth is generated by single models.

4.9. Conclusions

This chapter blazes the trail for the application of the model-averaging technique not only to volatility density as investigated here but to any density-forecasting aspiration. The spectrum here is twofold. First, the real-world predictive density under consideration is approximated using GARCH-type bootstrap methods. In the second part, the risk-neutral volatility density is scrutinized as this is derived using options written on the implied volatility index. In both cases averaging schemes outperformed single-model alternatives, substantiating in this way the comparative advantage of the method. In particular, Thick Model Averaging is the strongest candidate alternative under all various specification of the method (retention percentages and distributional assumptions of the volatility). In general, however, the gamma density is the weakest performing distributional pattern, while inferences of the volatility density under the empirical or the Log-Normal spectrum work better. An investment strategy where a long-short strategy is implemented concludes the chapter and highlights once more the superiority of the averaging schemes as opposed to single modelling alternatives.

[15] Tables 4.15, 4.16 include the collective rankings of all models estimated.

	Empirical Distribution		Log-Normal		Gamma	
Rank	**Volatility Density Model**	**P-value**	**Volatility Density Model**	**P-value**	**Volatility Density Model**	**P-value**
1	TM 40% AIC	0.093	TM 40% AIC	0.103	TM 40% AIC	0.038
2	TM 40% BIC	0.093	TM 40% BIC	0.103	TM 40% BIC	0.038
3	TM 30% AIC	0.084	TM 50% AIC	0.103	TM 50% AIC	0.036
4	TM 30% BIC	0.084	TM 50% BIC	0.103	TM 50% BIC	0.036
5	TM 50% AIC	0.084	TM 30% AIC	0.091	TM 30% AIC	0.002
6	TM 50% BIC	0.084	TM 30% BIC	0.091	TM 30% BIC	0.002
7	TM 20% AIC	0.056	TM 20% BIC	0.055	TM 20% AIC	0.002
8	TM 20% BIC	0.056	TM 20% AIC	0.055	TM 20% BIC	0.002
9	BA AIC	0.000	BA AIC	0.000	TM 10% AIC	0.001
10	BA BIC	0.000	BA BIC	0.000	TM 10% BIC	0.001
11	TM 10% AIC	0.000	TM 10% AIC	0.000	BA AIC	0.000
12	TM 10% BIC	0.000	TM 10% BIC	0.000	BA BIC	0.000
13	BEST AIC	0.000	BEST AIC	0.000	BEST AIC	0.000
14	BEST BIC	0.000	BEST BIC	0.000	BEST BIC	0.000

Table 4.5.: Rankings of the GARCH-type volatility density forecasting models according to the Empirical, Log-Normal and Gamma distribution.

Rank	Model	P-value Berkowitz	Rank	Model	P-value Berkowitz
1	TM-MSE 33% Log-Normal	0.7021	**19**	TM-MSE 50% Gamma	0.6074
2	TM-AIC 33% Log-Normal	0.6984	**20**	TM-BIC 33% Emprirical	0.5981
3	TM-BIC 50% Log-Normal	0.6897	**21**	BA-MSE Gamma	0.5841
4	TM-MSE 50% Log-Normal	0.6840	**22**	TM-BIC 33% Gamma	0.5822
5	TM-AIC 50% Log-Normal	0.6823	**23**	BA-AIC Gamma	0.5820
6	TM-MSE 50% Emprirical	0.6554	**24**	BA-BIC Gamma	0.5803
7	TM-BIC 50% Emprirical	0.6542	**25**	BA-MSE Empririca l	0.5750
8	TM-BIC 33% Log-Normal	0.6520	**26**	BA-AIC Empririca l	0.5729
9	TM-MSE 33% Emprirical	0.6500	**27**	BA-BIC Empririca l	0.5711
10	TM-AIC 33% Emprirical	0.6482	**28**	BEST MSE Log-Normal	0.5496
11	TM-AIC 50% Emprirical	0.6477	**29**	BEST MSE Emprirical	0.5240
12	BA-MSE Log-Normal	0.6354	**30**	BEST BIC Log-Normal	0.5105
13	BA-AIC Log-Normal	0.6332	**31**	BEST AIC Log-Normal	0.5087
14	BA-BIC Log-Normal	0.6313	**32**	BEST AIC Emprirical	0.4940
15	TM-MSE 33% Gamma	0.6268	**33**	BEST BIC Emprirical	0.4895
16	TM-AIC 33% Gamma	0.6246	**34**	BEST MSE Gamma	0.4718
17	TM-BIC 50% Gamma	0.6167	**35**	BEST BIC Gamma	0.4376
18	TM-AIC 50% Gamma	0.6081	**36**	BEST AIC Gamma	0.4327

Table 4.6.: Rankings of the risk-neutral volatility density forecasting models and averaging schemes.

Rank	Model	P-value Berkowitz	Rank	Model	P-value Berkowitz
1	TM-AIC 33% Emprirical	0.1047	**19**	BEST BIC Gamma	0.0770
2	TM-MSE 50% Emprirical	0.0985	**20**	BEST BIC Log-Normal	0.0720
3	TM-AIC 33% Log-Normal	0.0974	**21**	BEST BIC Emprirical	0.0741
4	TM-MSE 50% Log-Normal	0.0962	**22**	TM-AIC 50% Gamma	0.0761
5	TM-BIC 33% Emprirical	0.0910	**23**	TM-BIC 50% Gamma	0.0759
6	TM-MSE 33% Emprirical	0.0910	**24**	BEST MSE Gamma	0.0758
7	TM-AIC 50% Emprirical	0.0894	**25**	BEST MSE Log-Normal	0.0805
8	TM-MSE 33% Log-Normal	0.0890	**26**	BEST MSE Emprirical	0.0696
9	TM-AIC 50% Log-Normal	0.0880	**27**	BA-BIC Emprirical	0.0743
10	TM-BIC 50% Log-Normal	0.0877	**28**	BA-AIC Emprirical	0.0740
11	TM-BIC 50% Emprirical	0.0860	**29**	BA-MSE Emprirical	0.0737
12	TM-MSE 50% Gamma	0.0851	**30**	TM-MSE 33% Gamma	0.0728
13	TM-BIC 33% Log-Normal	0.0846	**31**	BA-BIC Log-Normal	0.0695
14	TM-AIC 33% Gamma	0.0840	**32**	BA-AIC Log-Normal	0.0690
15	BEST AIC Gamma	0.0803	**33**	BA-MSE Log-Normal	0.0685
16	BEST AIC Log-Normal	0.0801	**34**	BA-BIC Gamma	0.0677
17	BEST AIC Emprirical	0.0748	**35**	BA-AIC Gamma	0.0674
18	TM-BIC 33% Gamma	0.0780	**36**	BA-MSE Gamma	0.0670

Table 4.7.: Rankings of Objective volatility density forecasting models and averaging schemes.

Rank	Model	cum ret	Rank	Model	cum ret
1	BEST AIC Emprirical	1.284	21	TM-BIC 50% Gamma	1.284
2	BA-AIC Emprirical	1.284	22	BEST MSE Gamma	1.284
3	TM-AIC 33% Emprirical	1.284	23	BA-MSE Gamma	1.284
4	TM-AIC 50% Emprirical	1.284	24	TM-MSE 33% Gamma	1.284
5	BEST BIC Emprirical	1.284	25	TM-MSE 50% Gamma	1.284
6	BA-BIC Emprirical	1.284	26	BEST AIC Log-Normal	1.284
7	TM-BIC 33% Emprirical	1.284	27	BA-AIC Log-Normal	1.284
8	TM-BIC 50% Emprirical	1.284	28	TM-AIC 33% Log-Normal	1.284
9	BEST MSE Emprirical	1.284	29	TM-AIC 50% Log-Normal	1.284
10	BA-MSE Emprirical	1.284	30	BEST BIC Log-Normal	1.284
11	TM-MSE 33% Emprirical	1.284	31	BA-BIC Log-Normal	1.284
12	TM-MSE 50% Emprirical	1.284	32	TM-BIC 33% Log-Normal	1.284
13	BEST AIC Gamma	1.284	33	TM-BIC 50% Log-Normal	1.284
14	BA-AIC Gamma	1.284	34	BEST MSE Log-Normal	1.284
15	TM-AIC 33% Gamma	1.284	35	BA-MSE Log-Normal	1.284
16	TM-AIC 50% Gamma	1.284	36	TM-MSE 33% Log-Normal	1.284
17	BEST BIC Gamma	1.284	37	TM-MSE 50% Log-Normal	1.284
18	BA-BIC Gamma	1.284	38	S&P 500	1.041
19	TM-BIC 33% Gamma	1.284			

Table 4.8.: Cumulative wealth generated by the investment strategy based on risk-neutral volatility density forecasting models and averaging schemes.

Rank	Model	cum ret	Rank	Model	cum ret
1	BA-AIC Emprirical	1.812	20	TM-AIC 33% Gamma	1.284
2	BA-BIC Emprirical	1.812	21	TM-AIC 50% Gamma	1.284
3	BA-MSE Emprirical	1.812	22	BEST BIC Gamma	1.284
4	BA-AIC Gamma	1.812	23	TM-BIC 33% Gamma	1.284
5	BA-BIC Gamma	1.812	24	TM-BIC 50% Gamma	1.284
6	BA-MSE Gamma	1.812	25	BEST MSE Gamma	1.284
7	BA-AIC Log-Normal	1.812	26	TM-MSE 33% Gamma	1.284
8	BA-BIC Log-Normal	1.812	27	TM-MSE 50% Gamma	1.284
9	BA-MSE Log-Normal	1.812	28	BEST AIC Log-Normal	1.284
10	BEST AIC Emprirical	1.284	29	TM-AIC 33% Log-Normal	1.284
11	TM-AIC 33% Emprirical	1.284	30	TM-AIC 50% Log-Normal	1.284
12	TM-AIC 50% Emprirical	1.284	31	BEST BIC Log-Normal	1.284
13	BEST BIC Emprirical	1.284	32	TM-BIC 33% Log-Normal	1.284
14	TM-BIC 33% Emprirical	1.284	33	TM-BIC 50% Log-Normal	1.284
15	TM-BIC 50% Emprirical	1.284	34	BEST MSE Log-Normal	1.284
16	BEST MSE Emprirical	1.284	35	TM-MSE 33% Log-Normal	1.284
17	TM-MSE 33% Emprirical	1.284	36	TM-MSE 50% Log-Normal	1.284
18	TM-MSE 50% Emprirical	1.284	37	S&P 500	1.041
19	BEST AIC Gamma	1.284			

Table 4.9.: Cumulative wealth generated by the investment strategy based on Objective volatility density forecasting models and averaging schemes.

Rank	Model	P-value Berkowitz	Rank	Model	P-value Berkowitz
1	JD Gamma	0.7367	**28**	BA-BIC Gamma	0.5803
2	SV-SW Gamma	0.7119	**29**	SV-SW Log-Normal	0.5791
3	TM-MSE 33% Log-Normal	0.7021	**30**	BA-MSE Emprirical	0.5750
4	TM-AIC 33% Log-Normal	0.6984	**31**	BA-AIC Emprirical	0.5729
5	TM-BIC 50% Log-Normal	0.6897	**32**	BA-BIC Emprirical	0.5711
6	TM-MSE 50% Log-Normal	0.6840	**33**	BEST MSE Log-Normal	0.5496
7	TM-AIC 50% Log-Normal	0.6823	**34**	BEST MSE Emprirical	0.5240
8	TM-MSE 50% Emprirical	0.6554	**35**	BEST BIC Log-Normal	0.5105
9	TM-BIC 50% Emprirical	0.6542	**36**	BEST AIC Log-Normal	0.5087
10	TM-BIC 33% Log-Normal	0.6520	**37**	BEST AIC Emprirical	0.4940
11	TM-MSE 33% Emprirical	0.6500	**38**	SV-SW Emprirical	0.4922
12	TM-AIC 33% Emprirical	0.6482	**39**	BEST BIC Emprirical	0.4895
13	TM-AIC 50% Emprirical	0.6477	**40**	BEST MSE Gamma	0.4718
14	BA-MSE Log-Normal	0.6354	**41**	DO Log-Normal	0.4666
15	BA-AIC Log-Normal	0.6332	**42**	DO Emprirical	0.4570
16	BA-BIC Log-Normal	0.6313	**43**	BEST BIC Gamma	0.4376
17	TM-MSE 33% Gamma	0.6268	**44**	BEST AIC Gamma	0.4327
18	TM-AIC 33% Gamma	0.6246	**45**	DO Gamma	0.4012
19	TM-BIC 50% Gamma	0.6167	**46**	LG Log-Normal	0.3971
20	JD Log-Normal	0.6098	**47**	LG Emprirical	0.3868
21	TM-AIC 50% Gamma	0.6081	**48**	LG Gamma	0.3838

22	TM-MSE 50% Gamma	0.6074	**49**	JD-CT Gamma	0.0433
23	TM-BIC 33% Emprirical	0.5981	**50**	JD-CT Log-Normal	0.0226
24	JD Emprirical	0.5862	**51**	JD-CT Emprirical	0.0018
25	BA-MSE Gamma	0.5841	**52**	SV-OUG Emprirical	0.0000
26	TM-BIC 33% Gamma	0.5822	**53**	SV-OUG Gamma	0.0000
27	BA-AIC Gamma	0.5820	**54**	SV-OUG Log-Normal	0.0000

Table 4.10.: Rankings of all risk-neutral volatility density forecasting models and averaging schemes.

Rank	Model	P-value Berkowitz	Rank	Model	P-value Berkowitz
1	TM-AIC 33% Emprirical	0.1047	**28**	BEST MSE Log-Normal	0.0805
2	LG Log-Normal	0.1037	**29**	BEST MSE Emprirical	0.0696
3	TM-MSE 50% Empririca l	0.0985	**30**	BA-BIC Empririca l	0.0743
4	TM-AIC 33% Log-Normal	0.0974	**31**	BA-AIC Emprirical	0.0740
5	TM-MSE 50% Log-Normal	0.0962	**32**	BA-MSE Emprirical	0.0737
6	TM-BIC 33% Emprirical	0.0910	**33**	TM-MSE 33% Gamma	0.0728
7	TM-MSE 33% Empririca l	0.0910	**34**	BA-BIC Log-Normal	0.0695
8	LG Gamma	0.0900	**35**	BA-AIC Log-Normal	0.0690
9	TM-AIC 50% Empririca l	0.0894	**36**	BA-MSE Log-Normal	0.0685
10	TM-MSE 33% Log-Normal	0.0890	**37**	DO Gamma	0.0685
11	LG Empririca l	0.0887	**38**	BA-BIC Gamma	0.0677
12	TM-AIC 50% Log-Normal	0.0880	**39**	BA-AIC Gamma	0.0674
13	TM-BIC 50% Log-Normal	0.0877	**40**	BA-MSE Gamma	0.0670
14	TM-BIC 50% Empririca l	0.0860	**41**	DO Empririca l	0.0668
15	TM-MSE 50% Gamma	0.0851	**42**	DO Log-Normal	0.0645
16	TM-BIC 33% Log-Normal	0.0846	**43**	JD Gamma	0.0420
17	TM-AIC 33% Gamma	0.0840	**44**	JD Log-Normal	0.0314
18	BEST AIC Gamma	0.0803	**45**	SV-SW Gamma	0.0200
19	BEST AIC Log-Normal	0.0801	**46**	SV-SW Log-Normal	0.0193
20	BEST AIC Empririca l	0.0748	**47**	JD Empririca l	0.0148
21	TM-BIC 33% Gamma	0.0780	**48**	JD-CT Gamma	0.0094

22	BEST BIC Gamma	0.0770	**49**	JD-CT Emprirical	0.0076
23	BEST BIC Log-Normal	0.0720	**50**	SV-OUG Gamma	0.0059
24	BEST BIC Emprirical	0.0741	**51**	JD-CT Log-Normal	0.0050
25	TM-AIC 50% Gamma	0.0761	**52**	SV-OUG Log-Normal	0.0016
26	TM-BIC 50% Gamma	0.0759	**53**	SV-SW Emprirical	0.0000
27	BEST MSE Gamma	0.0758	**54**	SV-OUG Emprirical	0.0000

Table 4.11.: Rankings of all Objective volatility density forecasting models and averaging schemes.

Rank	**Model**	**Distribution**	**P**	**Q**	**P-value**	**Rank**	**Model**	**Distribution**	**P**	**Q**	**P-value**
1	APGARCH	GED	4	4	0.779	**31**	APGARCH	GED	4	1	0.016
2	APGARCH	T	4	4	0.779	**32**	APGARCH	T	4	1	0.016
3	APGARCH	Normal	4	4	0.353	**33**	APGARCH	Skew-t	2	1	0.016
4	APGARCH	Skew-t	4	2	0.137	**34**	APGARCH	GED	2	3	0.014
5	TM 40% AIC	-	0	0	0.093	**35**	APGARCH	T	2	3	0.014
6	TM 40% BIC	-	0	0	0.093	**36**	APGARCH	GED	2	2	0.013
7	APGARCH	Normal	1	3	0.088	**37**	APGARCH	T	2	2	0.013
8	APGARCH	Skew-t	2	3	0.087	**38**	APGARCH	GED	1	2	0.013
9	TM 30% AIC	-	0	0	0.084	**39**	APGARCH	T	1	2	0.013
10	TM 30% BIC	-	0	0	0.084	**40**	APGARCH	Normal	2	2	0.013
11	TM 50% AIC	-	0	0	0.084	**41**	APGARCH	GED	1	1	0.012
12	TM 50% BIC	-	0	0	0.084	**42**	APGARCH	T	1	1	0.012

13	APGARCH	Normal	1	4	0.073	**43**	APGARCH	GED	2	4	0.011
14	APGARCH	Normal	4	3	0.057	**44**	APGARCH	T	2	4	0.011
15	APGARCH	Normal	1	2	0.056	**45**	APGARCH	GED	3	4	0.011
16	TM 20% AIC	-	0	0	0.056	**46**	APGARCH	T	3	4	0.011
17	TM 20% BIC	-	0	0	0.056	**47**	APGARCH	GED	3	3	0.011
18	APGARCH	Normal	1	1	0.054	**48**	APGARCH	T	3	3	0.011
19	APGARCH	Skew-t	3	2	0.047	**49**	APGARCH	Skew-t	3	3	0.011
20	APGARCH	Skew-t	2	4	0.041	**50**	APGARCH	GED	2	1	0.010
21	APGARCH	GED	4	3	0.039	**51**	APGARCH	T	2	1	0.010
22	APGARCH	T	4	3	0.039	**52**	APGARCH	Skew-t	3	4	0.008
23	APGARCH	GED	1	3	0.026	**53**	APGARCH	Normal	3	3	0.008
24	APGARCH	T	1	3	0.026	**54**	APGARCH	Normal	3	4	0.007
25	APGARCH	Skew-t	2	2	0.026	**55**	APGARCH	Normal	2	4	0.007
26	APGARCH	Normal	2	3	0.022	**56**	APGARCH	Normal	4	1	0.006
27	APGARCH	GED	1	4	0.018	**57**	APGARCH	Normal	3	1	0.006
28	APGARCH	T	1	4	0.018	**58**	APGARCH	Skew-t	4	1	0.006
29	APGARCH	GED	3	1	0.016	**59**	APGARCH	GED	4	2	0.006
30	APGARCH	T	3	1	0.016	**60**	APGARCH	T	4	2	0.006

Table 4.12.: Rankings of the first 60 out of 334 GARCH-type real-world volatility density forecasting models according to the empirical distribution.

Rank	Model	Distribution	P	Q	P-value	Rank	Model	Distribution	P	Q	P-value
1	APGARCH	Normal	2	3	0.208	**31**	TM 20% BIC	-	0	0	0.055
2	APGARCH	GED	2	2	0.184	**32**	TM 20% AIC	-	0	0	0.055
3	APGARCH	T	2	2	0.184	**33**	APGARCH	Normal	3	4	0.054
4	APGARCH	GED	4	1	0.156	**34**	APGARCH	Normal	1	4	0.051
5	APGARCH	T	4	1	0.156	**35**	APGARCH	Normal	2	4	0.037
6	APGARCH	GED	3	1	0.154	**36**	APGARCH	Normal	4	1	0.033
7	APGARCH	T	3	1	0.154	**37**	APGARCH	Normal	4	4	0.033
8	APGARCH	GED	2	3	0.153	**38**	APGARCH	Normal	3	1	0.029
9	APGARCH	T	2	3	0.153	**39**	APGARCH	GED	4	2	0.023
10	APGARCH	GED	3	4	0.131	**40**	APGARCH	T	4	2	0.023
11	APGARCH	T	3	4	0.131	**41**	APGARCH	Normal	3	2	0.019
12	APGARCH	GED	4	4	0.126	**42**	APGARCH	GED	3	2	0.019
13	APGARCH	T	4	4	0.126	**43**	APGARCH	T	3	2	0.019
14	APGARCH	GED	2	1	0.117	**44**	APGARCH	Normal	4	2	0.012
15	APGARCH	T	2	1	0.117	**45**	APGARCH	Normal	2	1	0.004
16	APGARCH	Normal	2	2	0.116	**46**	APGARCH	Skew-t	4	2	0.004
17	TM 40% AIC	-	0	0	0.103	**47**	GJRGARCH	Skew-t	3	3	0.003
18	TM 40% BIC	-	0	0	0.103	**48**	APGARCH	GED	4	3	0.002
19	TM 50% AIC	-	0	0	0.103	**49**	APGARCH	T	4	3	0.002
20	TM 50% BIC	-	0	0	0.103	**50**	APGARCH	Skew-t	2	3	0.001
21	APGARCH	GED	3	3	0.092	**51**	APGARCH	Skew-t	3	2	0.001

22	APGARCH	T	3	3	0.092	**52**	APGARCH	Normal	4	3	0.001
23	TM 30% AIC	-	0	0	0.091	**53**	APGARCH	GED	1	3	0.001
24	TM 30% BIC	-	0	0	0.091	**54**	APGARCH	T	1	3	0.001
25	APGARCH	Normal	3	3	0.087	**55**	APGARCH	GED	1	4	0.001
26	APGARCH	GED	2	4	0.080	**56**	APGARCH	T	1	4	0.001
27	APGARCH	T	2	4	0.080	**57**	APGARCH	Skew-t	2	4	0.001
28	APGARCH	Normal	1	3	0.063	**58**	APGARCH	Skew-t	1	4	0.001
29	APGARCH	Normal	1	1	0.059	**59**	GJRGARCH	Normal	3	3	0.001
30	APGARCH	Normal	1	2	0.055	**60**	APGARCH	Skew-t	2	2	0.001

Table 4.13.: Rankings of the first 60 out of 334 GARCH-type real-world volatility density forecasting models according to the Log-Normal distribution.

Rank	**Model**	**Distribution**	**P**	**Q**	**P-value**	**Rank**	**Model**	**Distribution**	**P**	**Q**	**P-value**
1	GJRGARCH	Skew-t	3	3	0.066	**31**	GJRGARCH	GED	3	2	0.000
2	TM 40% AIC	-	0	0	0.038	**32**	GJRGARCH	T	3	2	0.000
3	TM 40% BIC	-	0	0	0.038	**33**	GARCH	Skew-t	3	4	0.000
4	TM 50% AIC	-	0	0	0.036	**34**	TGARCH	Normal	4	2	0.000
5	TM 50% BIC	-	0	0	0.036	**35**	GARCH	Skew-t	4	3	0.000
6	GJRGARCH	Normal	3	3	0.016	**36**	GJRGARCH	GED	4	3	0.000
7	GJRGARCH	Skew-t	3	4	0.016	**37**	GJRGARCH	T	4	3	0.000
8	GJRGARCH	Skew-t	4	3	0.009	**38**	GJRGARCH	Skew-t	3	1	0.000

9	GJRGARCH	Normal	3	4	0.009	**39**	TGARCH	GED	4	2	0.000
10	GJRGARCH	GED	3	3	0.005	**40**	TGARCH	T	4	2	0.000
11	GJRGARCH	T	3	3	0.005	**41**	GJRGARCH	GED	4	2	0.000
12	GJRGARCH	GED	3	4	0.002	**42**	GJRGARCH	T	4	2	0.000
13	GJRGARCH	T	3	4	0.002	**43**	TGARCH	GED	3	3	0.000
14	TM 30% AIC	-	0	0	0.002	**44**	TGARCH	T	3	3	0.000
15	TM 30% BIC	-	0	0	0.002	**45**	GJRGARCH	Skew-t	4	2	0.000
16	TM 20% AIC	-	0	0	0.002	**46**	GJRGARCH	Skew-t	4	4	0.000
17	TM 20% BIC	-	0	0	0.002	**47**	TGARCH	Normal	3	4	0.000
18	GARCH	Skew-t	3	3	0.001	**48**	TGARCH	Normal	4	3	0.000
19	GJRGARCH	Skew-t	3	2	0.001	**49**	TGARCH	Skew-t	3	2	0.000
20	TM 10% AIC	-	0	0	0.001	**50**	TGARCH	Normal	2	3	0.000
21	TM 10% BIC	-	0	0	0.001	**51**	TGARCH	GED	4	3	0.000
22	BA AIC	-	0	0	0.000	**52**	TGARCH	T	4	3	0.000
23	BA BIC	-	0	0	0.000	**53**	GJRGARCH	Normal	3	2	0.000
24	TGARCH	Normal	3	3	0.000	**54**	GJRGARCH	Normal	4	3	0.000
25	TGARCH	Normal	2	2	0.000	**55**	TGARCH	GED	2	3	0.000
26	TGARCH	Normal	3	2	0.000	**56**	TGARCH	T	2	3	0.000
27	TGARCH	GED	3	2	0.000	**57**	GJRGARCH	Skew-t	2	1	0.000
28	TGARCH	T	3	2	0.000	**58**	GJRGARCH	Normal	4	2	0.000
29	TGARCH	GED	2	2	0.000	**59**	GJRGARCH	GED	4	4	0.000

30	TGARCH	T	2	2	0.000	**60**	GJRGARCH	T	4	4	0.000

Table 4.14.: Rankings of the first 60 out of 334 GARCH-type real-world volatility density forecasting models according to the Gamma distribution.

Rank	**Model**	**cum ret**	**Rank**	**Model**	**cum ret**
1	LG Emprirical	1.284	**30**	LG Log-Normal	1.284
2	DO Emprirical	1.284	**31**	DO Log-Normal	1.284
3	BEST AIC Emprirical	1.284	**32**	SV-SW Log-Normal	1.284
4	BA-AIC Emprirical	1.284	**33**	BEST AIC Log-Normal	1.284
5	TM-AIC 33% Emprirical	1.284	**34**	BA-AIC Log-Normal	1.284
6	TM-AIC 50% Emprirical	1.284	**35**	TM-AIC 33% Log-Normal	1.284
7	BEST BIC Emprirical	1.284	**36**	TM-AIC 50% Log-Normal	1.284
8	BA-BIC Emprirical	1.284	**37**	BEST BIC Log-Normal	1.284
9	TM-BIC 33% Emprirical	1.284	**38**	BA-BIC Log-Normal	1.284
10	TM-BIC 50% Emprirical	1.284	**39**	TM-BIC 33% Log-Normal	1.284
11	BEST MSE Emprirical	1.284	**40**	TM-BIC 50% Log-Normal	1.284
12	BA-MSE Emprirical	1.284	**41**	BEST MSE Log-Normal	1.284
13	TM-MSE 33% Emprirical	1.284	**42**	BA-MSE Log-Normal	1.284
14	TM-MSE 50% Emprirical	1.284	**43**	TM-MSE 33% Log-Normal	1.284
15	LG Gamma	1.284	**44**	TM-MSE 50% Log-Normal	1.284

16	DO Gamma	1.284	**45**	SV-SW Emprirical	1.283
17	SV-SW Gamma	1.284	**46**	JD Emprirical	1.073
18	BEST AIC Gamma	1.284	**47**	JD Gamma	1.073
19	BA-AIC Gamma	1.284	**48**	JD Log-Normal	1.073
20	TM-AIC 33% Gamma	1.284	**49**	S&P 500	1.041
21	TM-AIC 50% Gamma	1.284	**50**	S&P 500	1.041
22	BEST BIC Gamma	1.284	**51**	S&P 500	1.041
23	BA-BIC Gamma	1.284	**52**	SV-OUG Emprirical	0.944
24	TM-BIC 33% Gamma	1.284	**53**	SV-OUG Gamma	0.944
25	TM-BIC 50% Gamma	1.284	**54**	SV-OUG Log-Normal	0.944
26	BEST MSE Gamma	1.284	**55**	JD-CT Emprirical	0.929
27	BA-MSE Gamma	1.284	**56**	JD-CT Gamma	0.929
28	TM-MSE 33% Gamma	1.284	**57**	JD-CT Log-Normal	0.929
29	TM-MSE 50% Gamma	1.284			

Table 4.15.: Cumulative wealth generated by the investment strategy based on all risk-neutral volatility density forecasting models and averaging schemes.

Rank	Model	cum ret	Rank	Model	cum ret
1	BA-AIC Emprirical	1.812	**30**	TM-AIC 33% Gamma	1.284
2	BA-BIC Emprirical	1.812	**31**	TM-AIC 50% Gamma	1.284
3	BA-MSE Emprirical	1.812	**32**	BEST BIC Gamma	1.284
4	BA-AIC Gamma	1.812	**33**	TM-BIC 33% Gamma	1.284
5	BA-BIC Gamma	1.812	**34**	TM-BIC 50% Gamma	1.284
6	BA-MSE Gamma	1.812	**35**	BEST MSE Gamma	1.284
7	BA-AIC Log-Normal	1.812	**36**	TM-MSE 33% Gamma	1.284
8	BA-BIC Log-Normal	1.812	**37**	TM-MSE 50% Gamma	1.284
9	BA-MSE Log-Normal	1.812	**38**	LG Log-Normal	1.284
10	JD Emprirical	1.514	**39**	DO Log-Normal	1.284
11	JD-CT Emprirical	1.514	**40**	BEST AIC Log-Normal	1.284
12	JD Gamma	1.514	**41**	TM-AIC 33% Log-Normal	1.284
13	JD-CT Gamma	1.514	**42**	TM-AIC 50% Log-Normal	1.284
14	JD Log-Normal	1.514	**43**	BEST BIC Log-Normal	1.284
15	JD-CT Log-Normal	1.514	**44**	TM-BIC 33% Log-Normal	1.284
16	LG Emprirical	1.284	**45**	TM-BIC 50% Log-Normal	1.284
17	DO Emprirical	1.284	**46**	BEST MSE Log-Normal	1.284
18	BEST AIC Emprirical	1.284	**47**	TM-MSE 33% Log-Normal	1.284
19	TM-AIC 33% Emprirical	1.284	**48**	TM-MSE 50% Log-Normal	1.284
20	TM-AIC 50% Emprirical	1.284	**49**	SV-OUG Emprirical	1.264

21	BEST BIC Emprirical	1.284	**50**	SV-SW Log-Normal	1.264
22	TM-BIC 33% Emprirical	1.284	**51**	SV-SW Emprirical	1.198
23	TM-BIC 50% Emprirical	1.284	**52**	SV-SW Gamma	1.198
24	BEST MSE Emprirical	1.284	**53**	SV-OUG Gamma	1.073
25	TM-MSE 33% Emprirical	1.284	**54**	SV-OUG Log-Normal	1.073
26	TM-MSE 50% Emprirical	1.284	**55**	S&P 500	1.041
27	LG Gamma	1.284	**56**	S&P 500	1.041
28	DO Gamma	1.284	**57**	S&P 500	1.041
29	BEST AIC Gamma	1.284			

Table 4.16.: Cumulative wealth generated by the investment strategy based on all Objective volatility density forecasting models and averaging schemes.

Date	**Number of Options**	**Date**	**Number of Options**
Mar-06	7	Jan-08	61
Apr-06	13	Feb-08	70
May-06	22	Mar-08	55
Jun-06	21	Apr-08	65
Jul-06	15	May-08	55
Aug-06	19	Jun-08	51
Sep-06	31	Jul-08	69
Oct-06	28	Aug-08	43

Nov-06	36	Sep-08	54
Dec-06	48	Oct-08	58
Jan-07	43	Nov-08	64
Feb-07	39	Dec-08	48
Mar-07	60	Jan-09	60
Apr-07	36	Feb-09	50
May-07	35	Mar-09	71
Jun-07	56	Apr-09	48
Jul-07	44	May-09	40
Aug-07	83	Jun-09	58
Sep-07	71	Jul-09	72
Oct-07	56	Aug-09	62
Nov-07	65	Sep-09	77
Dec-07	40	Oct-09	83

Table 4.17.: Number of options on the VIX index used to calculate each surface.

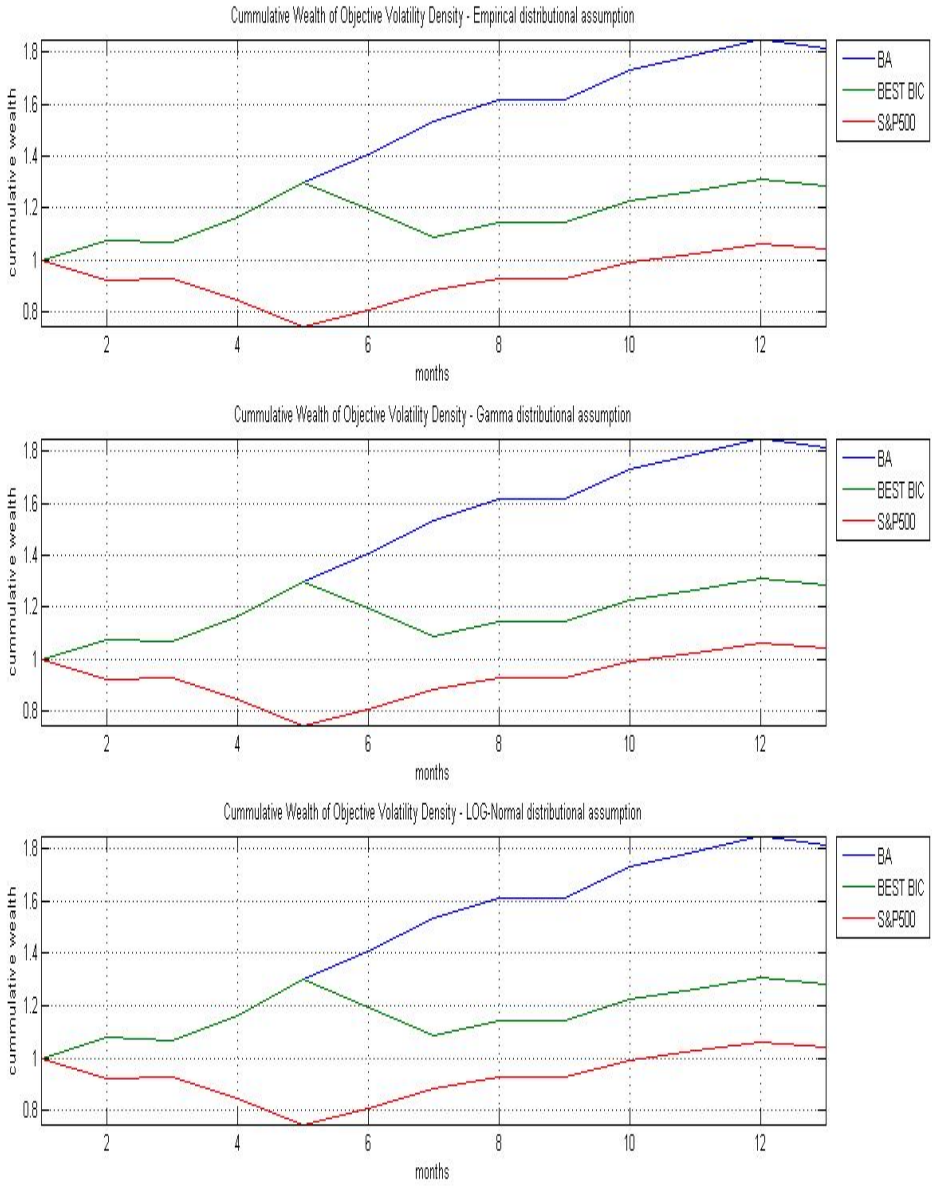


Figure 4.1.: Plot of the cumulative wealth generated by the investment strategy based on Objective volatility density forecasting processes under all three distributional assumptions and in particular for the Best single models selected according to BIC, for the Thick Model Averaging method using BIC and for the S&P500 index.

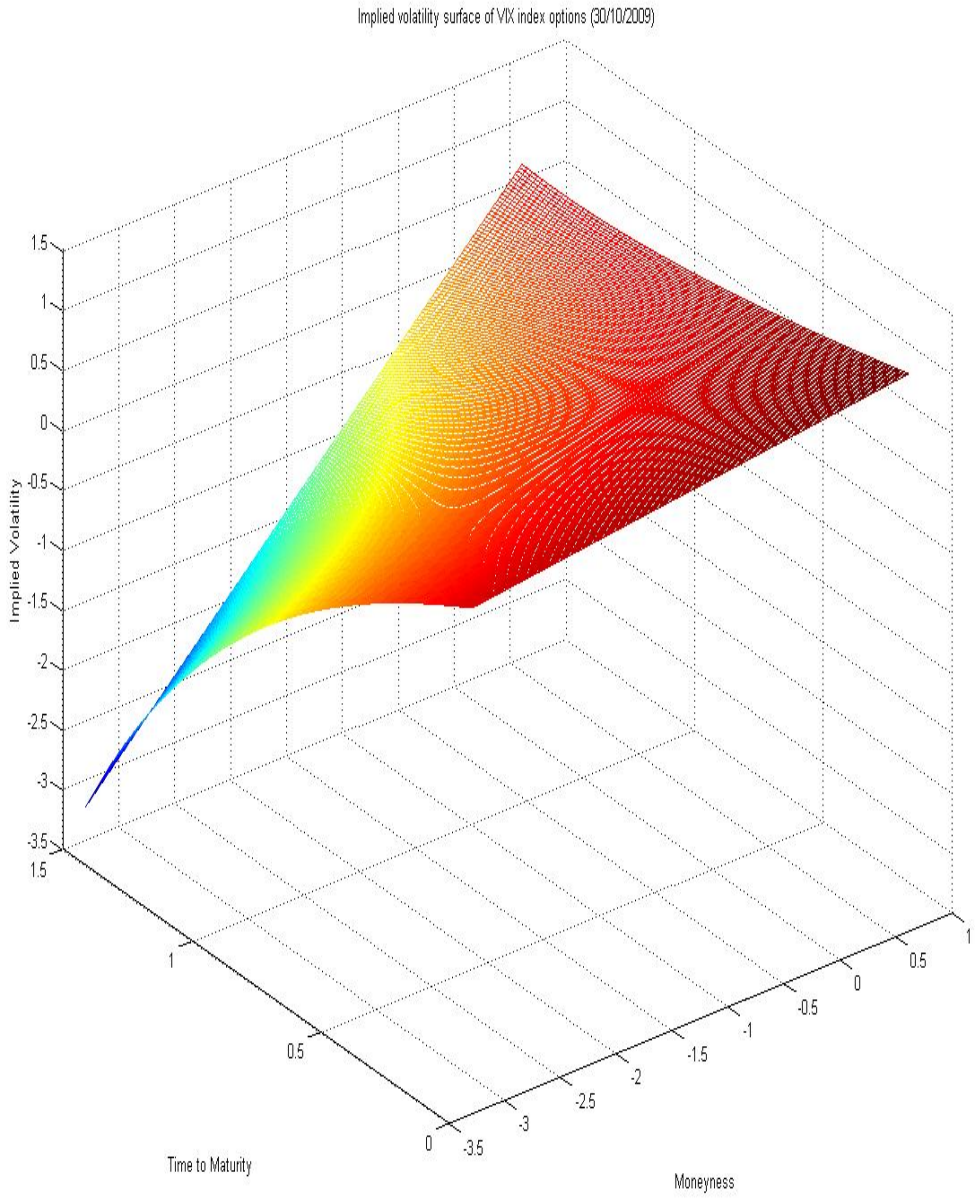


Figure 4.2.: Implied volatility surface based on options on the VIX index.

5. Variable selection and application in Asset allocation problems

5.1. Introduction

The question of whether asset predictability can actually be achieved and sustained for a relative period of time, or is just a random outcome produced sporadically from analysts that interminably try to 'beat the market' and generate excess profits, has yet to be answered. Fama (1970), based on his forerunners, Kendall (1953), Roberts (1959), Osborn (1959), Alexander (1961), Samuelson (1965) and Mandelbrot (1966) among others, argues that the efficient market hypothesis can be substantiated and offer tests to support the random-walk model assumption. This view was widely acknowledged until the mid eighties, when further indications seriously questioned its validity. Frictionless markets could for the first time be systematically exploited and the strong form of efficiency no longer fitted empirical predictive outcomes.

This evidence is mostly related to financial or macroeconomic indicators, along with corporate ratios as derived from financial statements, all publicly available. Some initial attempts to use historic information in order to foresee movements of the various financial instruments include Campbell and Shiller (1991), who try to predict interest rates using yield spreads, Fama and French (1988), who use dividend yields to forecast stock returns, and many others. More recent work not only comes as additional evidence of the catholic rejection of Fama's initial view of a non-predictable market, but also gives rise to the development of the field of Behavioural Finance, as well as to the "Technical Analysis" domain. Advocates of the first include Lo with his work in the "Adaptive Market Hypothesis", where he questions

the efficient market hypothesis (EMI), while empirical evidence is provided in Lo and MacKinlay (1999) and Lo *et al.* (2000). In addition, Shiller in his earlier work (1981) argues that stock prices relate to future dividends, while in a more contemporary paper (Shiller (1996)) he uses P/E ratios to forecast returns. Other influential researchers related to the behavioural finance discipline include (few among many) Shleifer (2000), Odean (1999), Schwert (2001) and Hirshleifer (1998, 2001).

If such an approach has been proved and accepted (Famma (1991), Pesaran and Timmermann (1995)) then a more serious problem unavoidably arises. That is to say how the investor can identify among a plethora of candidate predictors only those that can offer him superior results. In other words, when trying to achieve asset predictability, this reduces to a variable-candidate predictor choice dilemma.

To this ends in this chapter we try to exploit algorithmic techniques related to the variable-selection predicament. These methods belong to the linear estimators regime and include stepwise regression, Ridge regression, all-subsets regression, LASSO, LARS and more recently Elastic Net and Sparse Principal Components Analysis (SPCA).[1] Undeniably variations of the above methods do exist in an attempt to capture some of their internal flaws, but the above selection is based on the wider acceptance which they have received as successful variable-selection algorithms with universal applications to a variety classes of problems.

It is also notable that some of the predetermined methods used in the analysis date as back as 1960 (stepwise regression) while others are more recent (for example SPCA and LARSEN were only introduced in 2005). As a result, testing those algorithms well established for decades against other cutting-edge proposals can offer an additional dimension to the empirical research. Not only that, but also examining the performance of their combination when averaging is performed could provide an insight regarding the dynamic of all these models as forecasting vehicles.

Historically, variable selection algorithms have been used for identification purposes regarding the underlying process that can capture movements of the dependent variable more efficiently than the common multiple regression procedure. Critical views of the methods do exist, mainly referring to problems related to multicollinearity, i.e. when predictors are a linear trans-

[1]Further details regarding these algorithms will be provided later in the chapter.

formation of other predictors, thus they are correlated. This creates biased estimators as opposed to those produced by linear regression. However a small introduction of bias is counteracted by the reduction in the variance of the parameter estimates which as a result generates more reliable predictive models. Marquardt (1970) was among the first to point out the superiority of Ridge and general inverse estimators against least-squares regression. Overall it can be justified that multicollinearity is only a problem for larger dimensional datasets, since a small amount of correlation of predictors may even randomly appear when performing regression.

Notable is also the fact that although a larger number of complicated mathematical algorithms has been designed in order to identify every time the most significant predictors, little or even no evidence of the application of them in financial problems exists. Their use is merely restricted in domains such as Genetics, Bioengineering, Biomedical Science and other relevant scientific fields. A common example of their utilization is facilitating the identification process of those important genes that can reveal genetic predisposition to certain deceases.

In Finance, though, the ability to identify only this elite minority of predictors that can produce the optimal forecast in different scenarios, time horizons and under various restrictions has been a wishful thought for years. The common approach suggests selecting a model that incorporates only economic relevant regressors in order to predict excess returns and disregards all other possible model alternatives using standard model selection criteria (Information Criteria, R^2, Mean Squared Error etc.) or other financial criteria (see among others Pesaran and Timmermann (1995)).

This chapter, however, tries to stretch the barrier above and beyond the scientific areas in which variable selection algorithms have been used until now, and offer the financial investor a tool towards a safer prediction and ultimately a safer trading strategy. The focal point here, though, is twofold. While the problem of variable selection within an asset predictability framework is under scrutiny, model selection uncertainty is also being considered. First, by taking into account a pool of predictors and utilizing all those well-established algorithmic techniques that exist, we try to identify and select only those variables that have the most economic significance. Afterwards, we use them as a forecasting vehicle in order to achieve predictive supremacy.

Second, instead of variable selection the problem of model selection is under investigation, and we account for that by using model averaging. Two distinct techniques for averaging are adopted: Bayesian Approximation and Thick Modelling. A consideration is also made when model averaging is performed, since uncertainty can also be introduced among the steps on the path that every algorithm produces towards convergence. The optimal cutting point for the researcher is selected, again using standard model selection criteria which unavoidably create insecurity. This has as a consequence first to look at averaging individual steps on the path, so that path dependency of an individual candidate model is eliminated, and second to average across all paths that all models produce, so that model uncertainty is tackled as well.

The chapter is organized as follows. First the theoretical background surrounding some of the most important linear variable selection algorithms is introduced, along with a detailed description of their application in steps since they will be used in the empirical analysis that follows. Second, this restrictive linear prerequisite is relaxed and extended to a more general regression framework. In the section that follows the Logistic regression and the transformation of the previously discussed algorithms in order to be applied under this new framework becomes the focal point. This is due to the fact that as well as predicting excess returns, an investor is also interested in forecasting the direction of the movement.

Additionally, model averaging along with some background needed for the implementation of Thick Model Averaging and Bayesian Approximation under this new spectrum are discussed. The chapter concludes with the empirical analysis which targets not only achieving optimal forecastability of an Equity, Bond, Commodity and Foreign Exchange index, but also generating profits based on investment strategies under the logistic framework. A portfolio using the four indexes is also constructed and under the Markowitz mean-variance framework portfolio performance is also investigated, using results both from linear and logistic regression. We close with some concluding remarks highlighting the importance and ultimate superiority of averaging techniques against single models as these are indicated from the empirical results.

5.2. Literature Review of Variable Selection Algorithms

Variable selection has been the centre of many scientific studies, since it is classified by the vast majority of researchers in relevant fields as the most important and daunting task to perform accurately. The reason behind this consideration is obvious, but the solution unfortunately does not share the same simplicity of thought. The voluminous literature that exists considering the above-mentioned problem stands as a living proof of the importance and extensive work devoted to the subject.

First, Efroymson (1960) proposes what is today known as the *stepwise regression*, a procedure that adds variables according to a statistic threshold. This can be considered as the great predecessor of all variable selection techniques and with three main variations. The first of these is the *forward stepwise regression*, where one begins with no variables and enters one by one as these surpass a threshold. As a second variation we find *backward stepwise regression*, where all available variables are considered at the beginning and afterwards they are discarded if they are at the bottom of the statistic's rankings and finally as the last alternative we have the *combination of the two (forward and backward)*. Although stepwise regression algorithm was a novel idea when first introduced, soon a critical consideration of the results it produced led many to dought its ability to generate unbiased coefficients, and most importantly, meaningful results. The flaw was in the pattern followed, where even if a predictor was non-optimal alone, it was excluded without taking into account its performance when combined with other candidate predictors.

The second well-known technique for variable selection is *all-subsets regression* (Furnival and Willson (1970)), which effectively considers all possible combinations of predictors in order to identify those that yield the best model. The apparent advantage of the procedure is that it does not isolate predictors in order to consider performances. Decisions that will eventually form a model do not depend on previous choices to exclude or include parameters. All $2^n - 1$ (where n is the number of parameters) alternative models are considered simultaneously and the one that produces the optimal fit is selected. However, not only was this method inadequate to effectively tackle some of the fundamental problems of stepwise regression, it also cre-

ates an even greater computational time burden. In other words, the major impediment of the algorithm lies in the need to regress as many as $2^n - 1$ times in the search for the best model and this figure grows exponentially in the number of predictors.

Although the above-mentioned algorithms (stepwise regression and all-subsets regression) account for the variable-selection difficulty, they do not address the problem of parameter estimation. On the contrary, *Ridge regression* (Draper and Smith (1998)) does the opposite. The method estimates parameters smaller though than the equivalent values calculated with ordinary least-squares regression. Under this regime the unnecessary explosion of the coefficients is avoided, though retaining the dimension of the problem, since all candidate variables remain in the prediction model. In general the parameters are estimated by minimizing the following equation.

$$\beta_{\substack{reridge \\ regression}} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{5.1}$$

where λ is the regularization parameter.

In the following decades, little progress was reported regarding alternative algorithms that could offer a more coherent procedure for addressing the problems of variable selection and parameter estimation successfully. Tibshirani (1996) was the first to make a step forward towards this direction. Not only that, but he also paved the way towards an unprecedented research interest for more advanced and efficient algorithmic alternatives. His model, *LASSO*, uses the standard linear regression practice where parameters are estimated by minimizing the residual sum of squares, but he adds an extra restriction to them such that the sum of their absolute values is less than or equal to a constant. This restriction has the property of shrinking some of the coefficients towards zero while retaining the rest in a sparse model. The minimization problem here takes the following form:

$$\beta_{LASSO} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{5.2}$$

Efron *et al.* (2004) proposes the *Least Angle Regression* algorithm (*LARS)* which relates to *stagewise* regression but uses geometric properties in order to move in the right direction when estimating parameters and faster.

LARS is considered the most celebrated model-selection proposition, and very rapidly attracted great interest from the scientific community. However, *LASSO* received equal attention and numerous extensions.[2] These include the *group LASSO* (Yuan and Lin, 2006, Zhao *et al.* 2008), where the predictors are separated in groups such that for each predictor a particular coefficient vector can be related. Groups as a whole can be excluded from or included into the model under this algorithm while the sparsity of *LASSO* is also retained here. Let us define first the *elliptical norm* as follows:

$$\|\eta\|_K = \left(n' K n\right)^{1/2} \quad \eta \in R^d, d \geq 1, \tag{5.3}$$

where $\underset{d\times d}{K}$ is a positive definite matrix. Then the minimization target to estimate the parameters is

$$\beta_{\substack{group \\ LASSO}} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \|\beta_j\|_{K_j}. \tag{5.4}$$

In this direction the *group LARS* (Yuan and Lin, 2006, Park and Hastie, 2006) has been also suggested. A recent addition to the family of variable selection algorithms is *Elastic Net (LARSEN)* by Zou and Hastie (2005), which targets datasets where the number of predictors is greater than the number of observations ($p >> n$), an attribute that LASSO cannot address effectively for this type of data sets. *LARSEN* applies both an ℓ_1 and an ℓ_2 norm penalty to the coefficients, thus creating "a grouping effect" and ultimately superior results against *LASSO*. This is formulated mathematically as

$$\beta_{LARSEN} = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2. \tag{5.5}$$

Another proposal directly related to issues where $p >> n$ *fused LASSO* (Tibshirani *et al.*, 2005). The algorithm can be applied mainly when there exists an ordering of the predictors, and it smooths the estimates by ap-

[2]For a more detailed review a good reference is: Hesterberg *et al.* (2008). Here we will refer to the most noteworthy extensions only.

plying a penalty both to the difference between successive coefficients and to parameters themselves. On the other hand, *smoothed LASSO* (Meir and Buhlmann, 2007) addresses the issue of adjacency related not to parameters this time but to data. In other words, when there exists an order in the observations, then one can account for this information by assigning a higher weight to those points. The coefficients are estimated by minimizing the following equation:

$$\beta_{\substack{fused \\ LASSO}} = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|. \quad (5.6)$$

Additionally some more innovative approaches have been introduced such as Sparse Principal Components Analysis (Zou, Hastie and Tibshirani, 2005).

Finally, there are some variations of LASSO penalty term and the inherent problems this creates. These variations include the *adaptive LASSO* (Zou, 2006), the *relaxed LASSO* (Meinshausen, 2007), and finally the *Dantzig selector* (Candes and Tao, 2007). The *relaxed LASSO* is a two-stage procedure, where first a LASSO estimation is performed and then for every model identified on the path a second *LASSO* regression is executed, for which however a different penalty term is applied, $\phi\lambda$, where $\phi \in [0, 1]$ and $\lambda \in [0, \infty,)$ with λ being the regular *LASSO* penalty responsible for assuring the sparsity of the model, while ϕ relates to the shrinkage of the coefficients, which is now smaller. To be more precise, the problem is formulated as follows:

$$\beta^{\phi,\lambda}{}_{\substack{relaxed \\ LASSO}} = \arg\min_{\beta} N^{-1} \sum_{i=1}^{N} \left(y_i - \sum_{j \in M_\lambda} \beta_j x_{ij} \right)^2 + \phi\lambda \left\| \beta \right\|_1, \quad (5.7)$$
$$\phi \in [0, 1] \ \textit{and} \ \lambda \in [0, \infty),$$

where M_λ is the set including the selected parameters which will be used in the estimation part. Finally, the *relaxed LASSO* is empirically proved to work better for larger-dimensional data sets and to offer superior performance to *LASSO*. A two-step alternative proposal is the *adaptive LASSO*, where first we use a consistent parameter estimator such as the least-squares

fit, and then we perform a *LASSO* regression where, as indicated from the first step, statistically important parameters are given less penalty-shrinkage. In mathematical notation the problem is formulated as below:

$$\hat{\beta}^{*(n)}{}_{\substack{adaptive \\ LASSO}} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \, |\beta_j|, \qquad (5.8)$$

where $\hat{\beta}$ is in general a $\sqrt{n}$-consistent estimator of β^* such as the OLS parameter vector estimate, β_{OLS}, and $\hat{w} = 1/\left|\hat{\beta}\right|^{\gamma}$ is the vector of weights with $\gamma > 0$.

Finally, the Dantzig selector is used for problems related to a greater number of predictors than observations ($n << p$), and can be formulated as a solution to an ℓ_1-norm minimization problem. To be more precise, the problem can be described as below.

Let $y = X\beta + z$ be the dependent variable expressed as a linear function of the independent matrix of predictors X with $z \overset{iid}{\sim} N\left(0, \sigma^2\right)$. Then the parameter estimate problem can be stated as

$$\min \left\| \tilde{\beta} \right\|_{\ell_1} \quad \text{subject to} \quad \|Xr\|_{\ell_\infty} \leq \left(1 + \frac{1}{t}\right) \sqrt{2 \log p} \sigma, \qquad (5.9)$$

where $r = y - X\tilde{\beta}$ is the error term vector and $t > 0$ is a scalar. The algorithm has not received many advocates since in lot of the cases LASSO seems to produce superior results. However, it retains its power in problems where $n << p$.

5.3. Theoretical Framework

5.3.1. The Linear Variable Selection Algorithms

LASSO (Tibshirani (1996)) is so renowned that can be clearly considered as the foundation on which all contemporary variable-selection algorithms were build. Everything that has been suggested afterwards is an extension or modification of this proposition. Tibishirani utilizes the widely acknowledged OLS framework to actually produce sparse models by restricting the parameters value below a threshold.

Initially, let us define X as a matrix of n observations and p predictors i.e $X = (x_1, x_2, \ldots, x_n)'$, $x_i = (1, x_{i2}, \ldots, x_{ip})'$ and $Y = (y_1, y_2, \ldots, y_n)'$ the dependent variable, and let us also impose that the X matrix is standardized and the Y vector is centered; then the *LASSO* estimate of the parameter vector β is

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \tag{5.10}$$

where $t \geq 0$ is a tuning parameter. This can also be expressed as a loss function (penalised residual sum of squares):

$$PRSS\left(\hat{\beta}\right) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \tag{5.11}$$

The second part of the expression, or the restriction imposed on the parameters is, also called ℓ_1-norm *penalty*.

While *LASSO* is used not only to estimate parameters but also for model selection purposes, *Ridge Regression* on the other hand does not create sparse models. It is used mainly to avoid undesirable explosion of parameter values, since if they are left unrestricted as in the OLS regression this can be inevitable. In other words, all variables are included using this method, but the values assigned are smaller than those in the regular ordinary least squares. This penalization-regularization of the coefficients tends to keep their variance low even when the model grows large.

In accordance with the previous notation for this method the parameters are estimated as follows:

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq t, \tag{5.12}$$

The equivalent loss function is

$$PRSS\left(\hat{\beta}\right) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 . \tag{5.13}$$

This ridge restriction in turn is called the ℓ_2-norm penalty and, in the majority of cases, does not shrink the parameters towards zero as the ℓ_1 *penalty* does. Since this is also a convex function, a unique global minimum exists. Taking first-order derivatives according to β, this yields the following solution:

$$\begin{aligned} \frac{\partial PRSS}{\partial\beta} &= -2X'(Y - X\beta) + 2\lambda\beta = 0, \\ \beta &= (X'X + \lambda I)^{-1} X'Y. \end{aligned} \tag{5.14}$$

The *LASSO* algorithm on the other hand does not have an analytical solution, and the researcher has to refer to the quadratic optimisation alternatives to assist the estimation. It is not until 2004 that Efron *et al.* introduced a method to efficiently account for some inherent inefficiencies of *LASSO*, utilizing algebraic properties that help the algorithm to reach convergence quickly. This is the Least Angle Regression for LASSO model, known as LARS. The algorithm in general can claim many attractive features (C_p statistic for selecting the optimal model on the path etc.) and for that reason it has attracted significant attention from the start.

Before referring to *LARS*, however, as the most acknowledged continuation of the *LASSO* it is preferable to mention another technique, the *Forward Stagewise Regression*, which is very much interrelated to *LARS*. The idea is identical with that used in the Forward Stepwise Regression, but with a difference. One by one, coefficients are indeed added to the model, but instead of moving from zero to the OLS estimate, the value of the parameter changes only by a small amount. To facilitate better understanding of the mechanics behind *LARS*, below we provide in steps the implementation of the algorithm.

- Firstly, all parameters are set equal to zero and the error vector equal to the response variable Y, which is centered.

- Secondly, the variable most correlated to the residuals is identified and the value of the coefficient related to that variable is increased by an amount δ, where $\delta_j = \varepsilon \cdot sign\left(corr\left(r, x_j\right)\right)$ with r the residual vector, x_j the variable with the highest correlation and ε a small initial value.

- Thirdly, the residual vector is renewed, $r \leftarrow r - \delta_j x_j$, again the variable with the highest correlation is identified and the relevant parameter is updated, $\beta_j \leftarrow \beta_j + \delta$, until no predictor is still correlated with the residuals.

Taking under consideration all the above about forward stagewise regression and *LASSO*, Least Angle Regression algorithm (or else '*LARS*') as proposed by Efron *et al.* in 2004 shares common ground of intuition. The algorithm starts as the forward stagewise regression, setting all parameters equal to zero and identifying the predictor variable that has the highest correlation with the dependent variable, which at step zero equals the residual vector. The coefficient related to this predictor then enters the model, leaving its zero starting point and moving towards its least-squares value until a second variable has as much correlation with the residual vector as the previous one. Contrary to stagewise regression, the algorithm moves from one point to another not in small steps but directly to the targeted value. This is achieved by progressing in an equiangular direction between the two predictors until a third one enters the model exhibiting the same correlation with the current residual vector as that of the previous two predictors. This continues until all predictors have entered the model.

More precisely, first we initialize all parameters to equal zero. Let $\hat{Y}$ be the dependent variable estimate, i.e. $\hat{Y} = X\beta$ with X, β as defined above. Then the initial residual vector equals $r = Y$ i.e. $\hat{Y}_0 = 0$, and in all subsequent steps $r = Y - \hat{Y}$. Secondly, we estimate the current correlation vector $\hat{c}$, when

$$\hat{c} = X'\left(Y - \hat{Y}\right), \tag{5.15}$$

and keep track of the covariates with the highest correlation,

$$\hat{C} = \max_j \left\{|\hat{c}_j|\right\}. \tag{5.16}$$

Let us also define an active subset A which will include the j position of the highest correlated variables,

$$A = \left\{ j : |\hat{c}_j| = \hat{C} \right\}, \tag{5.17}$$

and a vector s to get the sign of the correlation,

$$s = sign\left\{ \hat{c}_j \right\}, \tag{5.18}$$

where $j \in A$. The next step is to estimate the weights w needed to be assigned to each variable selected in order to proceed in an equiangular direction:

$$\begin{aligned} w_A &= A_A G_A^{-1} 1_A, \\ G_A &= X_A^{'} X_A \text{ (the Gram matrix)}, \\ X_A &= (\ldots s_j x_j \ldots)_{j \in A}, \\ A_A &= \left(1_A^{'} G_A^{-1} 1_A \right)^{-\frac{1}{2}}. \end{aligned} \tag{5.19}$$

The equiangular vector is simply the product of the weights with the active predictors,

$$u_A = X_A w_A. \tag{5.20}$$

Additionally we have to estimate the correlation α between all variables and the equiangular vector, which implies

$$\alpha \equiv X' u_A, \tag{5.21}$$

and at the end to update the Y_A as follows:

$$Y_{A+} = Y_A + \hat{\gamma} u_A, \tag{5.22}$$

$$\hat{\gamma} = \underset{j \in A^c}{\overset{+}{\min}} \left\{ \frac{\hat{C} - \hat{c}_j}{A_A - \alpha_j}, \frac{\hat{C} + \hat{c}_j}{A_A + \alpha_j} \right\}. \tag{5.23}$$

The '$+$' symbol over *min* implies that the terms in the curly brackets that are considered are only the positive ones. Finally, the beta coefficients are updated such that

$$\beta_{k+1} = \beta_k + \hat{\gamma} w_A^{'}, \tag{5.24}$$

where k is the step. Afterwards a variable is added and the procedure continues until all p predictors have entered the model.

After all possible estimates of the β vector are generated, the researcher has to select just a single candidate model. Efron *et al.* propose the well-known C_p criterion, an additional important contribution of the *LARS* algorithm since it provides the researcher a convenient and easily applicable tool to use as a model selection indicator.

Under the assumption that the response variable Y is distributed with mean μ and σ^2 variance, i.e. $Y \sim (\mu, \sigma^2 I)$, the C_p risk measure statistic is

$$C_p = \frac{\left\| Y - \hat{Y} \right\|^2}{\sigma^2} - \eta + 2df_{\mu,\sigma^2}, \tag{5.25}$$

where $df_{\mu,\sigma^2} = \sum_{i=1}^{n} \operatorname{cov}(\hat{y}_i, y_i)/\sigma^2$ and σ^2 the residuals variance, or simply for the *LARS* case an approximation yields $df_{\mu,\sigma^2} = k$ where k is the number of steps.[3] Other more general approaches for model selection can also be applied as k-fold cross-validation, which we propose here, or the selection of an information criterion such as the Bayesian or Akaike Information Criterion, or any other error measure (for example Mean Square Error etc.). All performances of these criteria will be evaluated later in this chapter and some conclusions will be discussed.

As an important alternative to the *LASSO* structure we find *Elastic Net*. It was initially suggested by Zou and Hastie in 2005 for dealing with data sets for which the number of predictors is greater than the number of observations, but turned out to produce in a many instances better results than its competing alternatives. First the *Naïve Elastic Net* is introduced, which includes both an l_1 and an l_2 norm penalty, i.e.:

$$\begin{aligned} PRSS(\beta) &= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \\ &= \|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2, \end{aligned} \tag{5.26}$$

[3] For a more complete reference and proof of why see the original paper of Efron *et al.* (2004).

or else setting $\alpha = \lambda_2/\lambda_1 + \lambda_2$ then we have to minimize the following to get the parameter estimates:

$$\begin{aligned} \beta &= \arg\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ \text{subject to} \qquad & (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t. \end{aligned} \tag{5.27}$$

The last term is called the elastic net penalty. It is also obvious that by setting $\alpha = 0$ we get a *LASSO* type optimisation problem, while setting $\alpha = 1$ we get a *Ridge Regression* equivalent. Zou and Hastie for this reason restrict the values of α such that $\alpha \in [0, 1]$. They also move a step forward in order to solve the optimisation problem of the *Naïve Elastic Net* and transform it into an equivalent "*LASSO*-type" problem.[4] Although the *Naïve Elastic Net* accounts for some of the already mentioned inefficiencies of *LASSO* and *Ridge Regression* algorithm, it is still an incompetent procedure, mainly because the two-fold penalty applied reduces the value of the coefficients twice, thus creating an additional fragility of parameter estimates. This problem forms the basis of the thought to use a correction term which in turn produces the so called Elastic Net algorithm (*LARSEN*).

We define $X^* = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix}$, $Y^* = \begin{pmatrix} Y \\ 0 \end{pmatrix}$ and $\beta^* = \sqrt{1+\lambda_2}\beta$ then the *Naïve Elastic Net* can be interpreted as "*LASSO*-type" problem as follows:

$$\beta^* = \arg\min \|Y^* - X^*\beta^*\|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \|\beta^*\|_1 . \tag{5.28}$$

Thus $\beta_{\substack{Naive \\ LARSEN}} = \frac{1}{\sqrt{1+\lambda_2}}\beta^*$. Now for the Elastic Net case we define $\beta_{LARSEN} = \sqrt{1+\lambda_2}\beta^*$ or else $\beta_{LARSEN} = (1+\lambda_2)\,\beta_{\substack{Naive \\ LARSEN}}$.

LARSEN has also applications in ***Principal Components Analysis*** as a technique to improve some of the most important shortcomings of the method. Zou, Hastie and Tibshirani (2005) propose *Sparse Principal Components Analysis*, targeted at reducing the number of principal components which traditionally combine all variables available linearly and thus produce

[4] For further details refer to Zou & Hastie (2006).

sparse and more interpretable models. The basic notion underpinning the model is to express the PCA problem in a regression context; We may thus derive the PCs easily through optimisation. Then we can straightforwardly use a variable-reduction method, here *LARSEN*, and produce the desired sparse PCs. The steps required can be described as follows.

First, suppose again that we have an X matrix that includes all variables to be used in the analysis. This matrix, following Singular Value Decomposition, can be expressed as

$$X = U\Sigma V', \tag{5.29}$$

where U is a matrix containing the PCs, Σ is a non-negative diagonal matrix containing the volatility of the PCs and finally V is a matrix containing the loadings. Since the PCs are a linear combination of the p variables the authors prove that we can recover the loadings by regressing the PCs on the p variables. We assume

$$\hat{\beta}_{ridge} = \arg\min \|Y_i - X\beta\|^2 + \lambda \|\beta\|^2 ;$$

then defining Y as the PC vector such that

$$Y = U\Sigma, \tag{5.30}$$

with Y_i being the ith PC, let the loadings vector $\hat{v} = \hat{\beta}_{ridge}/|\hat{\beta}_{ridge}|$, then

$$\hat{v} = V_i, \tag{5.31}$$

where V_i is the ith loading. This result, although important, is not a genuine alternative. We have first to perform PCA, and then to use (5.30) to find a suitable sparse approximation. This can only be regarded as our firs step towards a more general transformation that will allow us not only to rebuild the loadings based on the initial SVD estimates but also to provide a self-contained regression-type framework to derive the PCs.

$$\begin{aligned}(\alpha, \beta) &= \arg\min \sum_{i=1}^{n} |x_i - \alpha\beta' x_i|^2 + \lambda \sum_{j=1}^{k} |\beta_j|^2 \\ &= TrX'X - 2Tr(\alpha'X'X\beta) + Tr\beta'(X'X + \lambda)\beta\end{aligned}$$

$$= \quad TrX'X + \sum_{j=1}^{k} \left({\beta_j}'(X'X + \lambda)\beta_j - 2{\alpha_j}'X'X{\beta_j}' \right), \quad (5.32)$$

where k is the number of principal components desired and is selected subjectively by the researcher (usually five to seven components are enough, but that depends on the nature of the problem) and x_i has already been defined above as $x_i = (1, x_{i2}, \ldots, x_{ip})'$, as taken from $X = (x_1, x_2, \ldots, x_n)'$ where n is the number of observations.

Setting α constant β can be estimated as follows:

$$\begin{aligned} \frac{\partial PRSS}{\partial \beta} &= 2\beta(X'X + \lambda) - 2X'X\alpha = 0, \\ \beta &= (X'X + \lambda)^{-1}X'X\alpha. \end{aligned} \quad (5.33)$$

For α though we have the following maximization problem:

$$\begin{aligned} \hat{\alpha} &= \arg\max Tr\alpha' \left(X'X\right)\beta \\ &= \arg\max Tr\alpha' \left(X'X\right)\left(X'X + \lambda\right)^{-1}X'X\alpha \\ &= s_j V, \end{aligned} \quad (5.34)$$

where s is a sign vector such that $s = \begin{cases} 1 & \text{correlation} > 0 \\ -1 & \text{elsewhere} \end{cases}$.

This in turn produces $\hat{\beta} \propto V_i \quad \forall i = 1, 2, \ldots,$ or

$$\beta_j = s_j \frac{D_j^2}{D_j^2 + \lambda} V_j. \quad (5.35)$$

Adding the ℓ_1 norm criterion will automatically generate the desired lower-dimensional models,

$$\left(\alpha, \hat{\beta}\right) = \arg\min \sum_{i=1}^{n} \left|x_i - \alpha\beta' x_i\right|^2 + \lambda \sum_{j=1}^{k} \left|\beta_j\right|^2 + \sum_{j=1}^{k} \lambda_{1j} \left|\beta_j\right|_1, \quad (5.36)$$

with $\alpha'\alpha$ a unitary matrix, β_j proportional to V_j where $j = 1, \ldots, k$, and k the number of components. An additional observation is that the above equation represents the *Naïve Elastic Net*, which does not include

the regularization-correction term. But here this term plays a trivial role, since the acquired beta estimates are normalized so as to get the V_i loadings. Some final details regarding the numerical solution proposed are also important for building the algorithm. For all asset classes, we have

$$\begin{aligned}\left(\alpha,\hat{\beta}\right) &= \arg\min\sum_{i=1}^{n}\left|x_i-\alpha\beta' x_i\right|^2+\lambda\sum_{j=1}^{k}\left|\beta_j\right|^2+\sum_{j=1}^{k}\lambda_{1j}\left|\beta_j\right|_1 \\ &= TrX'X-2Tr\alpha'X'X\beta+Tr\beta'(X'X+\lambda)\beta+\sum_{j=1}^{k}\lambda_{1j}\left|\beta_j\right|_1 \\ &= TrX'X+\sum_{j=1}^{k}\left(\beta_j{}'(X'X+\lambda)\beta_j-2\alpha_j{}'X'X\beta_j{}'+\lambda_{1j}\left|\beta_j\right|_1\right)\end{aligned} \tag{5.37}$$

To solve for each parameter we should consider two distinct equations:[5]

$$\begin{aligned}\beta_j &= \arg\min(1+\lambda)\beta_j'\left(X'X+\lambda\right)(1+\lambda)\beta_j-\ldots \\ &- 2\alpha_j'X'X(1+\lambda)\beta_j+\lambda_{1j}\left|(1+\lambda)\beta_j\right|_1, \qquad (5.38)\\ \alpha_j &= \arg\min Tr\alpha_j'\left(X'X\right)\beta_j \\ &= U_jV_j. \qquad (5.39)\end{aligned}$$

Summarizing, the algorithm builds as follows.

- First ,we perform singular-value decomposition on our initial matrix X such that from this transformation three new matrices are created, U, Σ, V'. The V matrix represents the loadings whose value at start is assigned to alpha parameter, $\alpha_0 = V$.
- Next, we calculate the β matrix of parameters, where β is a $p \times k$ matrix such that β_j represents the output of the *LARSEN* algorithm and is $p \times 1$ vector where $j = 1, \ldots,\ k$ and k the number of components (as defined above). What feeds into the *LARSEN* algorithm is X as the independent variable and $X\alpha_j$ representing the dependent variable.

[5]For proof of that conclusion along with more detailed analysis refer to Zou *et al.* (2006).

- Afterwards we normalize the β output of the k - *LARSEN* algorithms such that their Euclidian distance equals 1 and check whether the minimization of β error estimate has been achieved and no further improvement can be accomplished. If we have no convergence we have to update alpha by performing singular-value decomposition on $X'X\beta$ taking new $U,\ \Sigma,\ V'$ and setting $\alpha_{new} = UV'$.

- Finally, we repeat the procedure until the optimal solution is found or the maximum number of iterations is reached.

From the analysis above it is evident that we are using only the X matrix, whose dimensions we try to reduce and as a result produce a sparse model that can be used subsequently. The most important output of SPCA for our analysis is the sparse loadings. The task now is to combine these with the initial X matrix (i.e. $X\beta_{SPCA}$) and obtain the sparse principal components ($X_{SPCA} = X\beta_{SPCA}$) which will be the new predictor variables. Afterwards we regress the SPCA on Y and compare the result $\left(\hat{Y} = X_{SPCA}\beta_{OLS}\right)$ with the true Y vector of responses, which is of course the ultimate objective in order to make forecasts.

5.3.2. The Generalized Linear Variable Selection Algorithms

Until now the discussion was primarily confined to the linear case, where the optimisation function only included linear combination of the parameters with the prediction variables. There are, however, several occasions where this is inefficient and a more general context should be used instead for the variable selection algorithms discussed above. This need created several extensions, all in an attempt to create robust optimisation routines that could provide a solution to the unavoidable problems that arise due to these higher-order relationships included in the optimisation functions.

If we rewrite the problem of variable selection in a more general setting, then we have the following equations:

Ridge-type Regression,

$$\hat{\beta} = \arg\min_{\beta} \left(L(\beta) + \lambda_2 \|\beta\|_2^2 \right); \tag{5.40}$$

LASSO-type Regression,

$$\hat{\beta} = \arg\min_{\beta} \left(L(\beta) + \lambda_1 \|\beta\|_1 \right); \tag{5.41}$$

Elastic Net-type Regression,

$$\hat{\beta} = \arg\min_{\beta} \left(L(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right); \tag{5.42}$$

where L is some general loss function. The focal point in this section though is the logistic loss function, as a great representative of the problem involved when dealing with binary data that can be described by a density function from the exponential family.

5.3.3. The Logistic Regression

Let us assume that the vector of true responses Y can take only two values,

$$Y = \begin{cases} 1 & Y > 0, \\ 0 & Y \leq 0, \end{cases} \tag{5.43}$$

with probability taken from the logistic distribution such that

$$\begin{aligned} \text{Prob}(Y = 1 \,|X) &= \frac{e^{\beta' X}}{1 + e^{\beta' X}}, & (5.44) \\ \text{Prob}(Y = 0 \,|X) &= 1 - \text{Prob}(Y = 1 \,|X) = \frac{1}{1 + e^{\beta' X}}. & (5.45) \end{aligned}$$

To estimate the parameters vector we must first define the likelihood function as

$$L(\beta \,|X) = \prod_{i=1}^{n} \text{Prob}(Y = 1 \,|X)^{Y_i} \left(1 - \text{Prob}(Y = 1 \,|X)\right)^{1 - Y_i}. \tag{5.46}$$

Taking logarithms on both sides the log-likelihood function is[6]

$$\begin{aligned}\log L(\beta) &= \sum_{i=1}^{n} y_i \log\left(\text{Prob}\left(Y=1\,|X=x_i\right)\right)+\ldots \\ &+ \quad (1-y_i)\log\left(1-\text{Prob}\left(Y=1\,|X=x_i\right)\right)\end{aligned}$$

or

$$\log L(\beta) = \sum_{i=1}^{n} y_i \beta' x_i - \log\left(1+e^{\beta' x_i}\right). \tag{5.47}$$

Since an analytical solution of parameters is not possible here, many approximations have been suggested instead. Amongst them, the Newton-Raphson optimisation algorithm received a considerable amount of attention and is more widely known as Iteratively Reweighted Least Squares (IRLS). However, the greatest drawback of the algorithm is the need to analytically calculate the Hessian matrix. Unfortunately, in most cases this is neither the easiest nor the safest path to follow. Briefly, the method can be described as follows. First, having in mind the above likelihood function for the logistic case, the parameter vector β is renewed using the formula

$$\beta_{new} = \beta_{old} - \left(\frac{\partial^2 L(\beta)}{\partial\beta\partial\beta'}\right)^{-1}\frac{\partial L(\beta)}{\partial\beta}, \tag{5.48}$$

$$\begin{aligned}\frac{\partial L(\beta)}{\partial\beta} &= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n}\frac{x_i e^{\beta' x_i}}{1+e^{\beta' x_i}} \\ &= \sum_{i=1}^{n} x_i\left(y_i - \frac{x_i e^{\beta' x_i}}{1+e^{\beta' x_i}}\right) \\ &= \sum_{i=1}^{n} x_i\left(y_i - \text{Prob}\left(x_i;\beta\right)\right), \\ \frac{\partial^2 L(\beta)}{\partial\beta\partial\beta'} &= -\sum_{i=1}^{n} x_i x_i^{'}\frac{e^{\beta' x_i}}{1+e^{\beta' x_i}}\left(1-\frac{e^{\beta' x_i}}{1+e^{\beta' x_i}}\right) \\ &= -\sum_{i=1}^{n} x_i \text{Prob}\left(x_i;\beta\right)\left(1-\text{Prob}\left(x_i;\beta\right)\right)x_i^{'}.\end{aligned}$$

[6]Taking into account the previous equation and by substitution we get this result.

Let

$$\begin{aligned}
\text{Prob}\,(x_i;\beta)\,(1-\text{Prob}\,(x_i;\beta)) &= P\,(1-P) = W, \\
\frac{\partial L(\beta)}{\partial \beta} &= X'\,(Y-P)\,, \\
\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} &= -X'WX.
\end{aligned}$$

Consequently

$$\beta_{new} = \beta_{old} + \left(X'WX\right)^{-1} X'\,(Y-P) \tag{5.49}$$

By multiplying β_{old} with $\left(X'WX\right)^{-1} X'WX$ and $\left(X'WX\right)^{-1} X'\,(Y-P)$ with WW^{-1}, we get

$$\begin{aligned}
\beta_{new} &= \left(X'WX\right)^{-1} X'WX\beta_{old} + \left(X'WX\right)^{-1} X'WW^{-1}\,(Y-P)\,, \\
\beta_{new} &= \left(X'WX\right)^{-1} X'W\left(X\beta_{old} + W^{-1}\,(Y-P)\right), \\
\beta_{new} &= \left(X'WX\right)^{-1} X'WZ, \qquad (5.50)
\end{aligned}$$

where $Z = X\beta_{old} + W^{-1}\,(Y-P)$ and is the so-called adjusted Y response. So the optimisation problem can be restated as

$$\beta_{new} \leftarrow \arg\min\,(Z-X\beta)'\,W\,(Z-X\beta)\,. \tag{5.51}$$

To summarise. the steps are as follows: First, the probability P and the diagonal matrix W are computed such that $\beta \leftarrow \left(X'WX\right)^{-1} X'WZ$. If the stopping criteria are met or the maximum number of iterations is reached, the algorithm stops. Otherwise a new probability P is computed along with a new W matrix, the β vector is updated and the stopping criteria are tested again until convergence is achieved.

The use of the Newton-Raphson algorithm is allowed here, since the logistic function that we want to estimate is concave, so a global minimum solution can be traced. But to facilitate the convergence of the algorithm a variation of the method has been proposed. This is the iterative 'while loop' by LeSagel (1999). This method depends on Newton's gradient idea for optimisation, where if the parameter vector β is not optimised a new value for that is calculated using a 'direction vector' Δ, where $\Delta = -H^{-1}\nabla_\beta L\,(\beta)$,

H^{-1} being the inverse of the Hessian matrix and $\nabla_\beta L(\beta)$ being the gradient with values as already calculated above. In other words the iteration loop continues using $\beta_{new} = \beta_{old} + \Delta$ until convergence is achieved.
In a more general context, the last equation can be rewritten as

$$\beta_{new} = \beta_{old} + \lambda\Delta, \tag{5.52}$$

where λ in Newton-Raphson's method is inferred to equal one. But LeSagel proposed λ, which can be regarded as a regularization of the size of the step Δ, also be computed iteratively. To be more precise, the procedure followed in order to determine the step size can be described as follows.

- First, the gradient vector and the Hessian matrix for the logistic model are estimated using the formulae mentioned above, i.e. given a response vector Y $\mathbf{1}_{Y>0}$ and a predictors matrix X the gradient equals

$$\nabla_{\beta_0} L(\beta_0) = X'(Y - P) \tag{5.53}$$

  and the Hessian matrix in turn equals

$$H = -X'WX, \tag{5.54}$$

  with

$$P = \frac{e^{\beta_0' x_i}}{1 + e^{\beta_0' x_i}}, \tag{5.55}$$

$$W = P(1 - P). \tag{5.56}$$

- Afterwards Δ_0 is calculated using for Δ_0 the formula

$$\Delta_0 = -H^{-1}\nabla_{\beta_0} L(\beta_0.) \tag{5.57}$$

- Then the determination of the optimal step λ is performed using a second smaller while-loop. An initial value of λ is first assigned (for example $\lambda_0 = 2$) and the logistic log-likelihood function using this λ is calculated,

$$\log L_1(\beta_1) = \sum_{i=1}^{n} y_i \beta_1' x_i - \log\left(1 + e^{\beta_1' x_i}\right), \tag{5.58}$$

where $\beta_1 = \beta_0 + \lambda_0 \Delta_0$, β_0 is a vector including starting values for the parameters and Δ_0 is the initial 'direction vector'.

- A recalculation of this logistic log-likelihood is performed using each time a new value for λ where $\lambda_{new} = \lambda_{old}/2$, until the new log-likelihood L_{new} is less than the previous log-likelihood, where

$$\log L_{new}\left(\beta_{new}\right) = \sum_{i=1}^{n} y_i \beta_{new}^{'} x_i - \log\left(1 + e^{\beta_{new}^{'} x_i}\right) \tag{5.59}$$

with

$$\beta_{new} = \beta_{old} + \frac{\lambda_{old}\Delta_{old}}{2}.$$

When this new likelihood takes a value less than that we already have it means that we have moved towards the right direction and accept the estimated value of λ.

- After λ and Δ, the parameter vector is now ready to be estimated as

$$\beta_{new} = \beta_{old} + \lambda_{old}\Delta_{old}.$$

- If convergence is not achieved, the update of the parameter vector should be performed again by changing the direction Δ along with the size of the step λ until a minimum is found.

5.3.4. ℓ_1 Regularized Logistic Regression

While the optimisation of the general logistic loss function is adroitly addressed by Newton methods, the additional constraint imposed on the parameters by an ℓ_1 type regularization term generates extra impediments that have been discussed extensively in the literature. This is because although the general logistic function is convex and continuously differentiable, by adding the ℓ_1 regularization term it loses this property of differentiability. This in turn dictates a different approach from the common one to the optimisation problem of the Logistic loss function. Some of the methods used to handle ℓ_1 optimisation problems include Sub-Gradient strategies, unconstrained optimisation approaches where the function of interest is replaced by one that is differentiable (a Log-Barrier, epsL1 etc), is added and

finally the constrained optimisation techniques where the ℓ_1 term is added as constraint to the algorithm.[7] One of the most successful representatives of the latter subcategory for solving relevant ℓ_1 regularization problems is the IRLS-*LARS* (Iterative Reweighted Least Squares for *LARS*, Lee *et al.* 2006). It was specifically designed though to handle only the Logistic Loss function and not other more general Loss functions. However, it received a lot of attention, as it efficiently approximates the Logistic function while restraining the values of the parameters according to the ℓ_1 penalization term.

Lately, though, Schmidt *et al.* (2007) propose a faster and more general approach than IRLS-*LARS* which very competently takes into advantage the 'Two Metric Approach' (Gafni, Bertsekas (1983)) and addresses the minimization of the ℓ_1 regularization loss function as a constrained optimisation problem. As a result the Projection ℓ_1 Method, as it is named after by the authors, can be applied for various loss functions while rapid convergence is guaranteed. Briefly the algorithm can be described as follows.

Firstly we have to reformulate the problem as a non-negative constrained optimization equivalent. The parameter vector is transformed into two non-negative components such that

$$\beta = \left[\begin{array}{rcl} \beta^{+} & = & \beta_{old} \cdot 1\left(\beta_{old} \geq 0\right) \\ \beta^{-} & = & -\beta_{old} \cdot 1\left(\beta_{old} < 0\right) \end{array} \right], \tag{5.60}$$

where β_{old} is a $k \times 1$ vector (k is the number of predictors) and β is a $2k \times 1$ vector which transforms the loss function into a differentiable one. The problem now can be restated as:

$$L\left(\beta^{+}, \beta^{-}\right) = f\left(\beta^{+} - \beta^{-}\right) + \lambda \sum_{i} \left(\beta_i^{+} + \beta_i^{-}\right) \quad \forall \beta_i^{+},\ \beta_i^{-} \geq 0, \tag{5.61}$$

where f here is the logistic log-likelihood as defined previously, but it can be replaced by other log-likelihood functions. Now that we rewritten the loss function, the gradient and the Hessian matrix can be estimated:

[7] For further details regarding available techniques for ℓ_1 optimisation problems refer to Schmidt *et al.* (2007).

$$\nabla_{\beta^+,\beta^-} L\left(\beta^+,\beta^-\right) = \begin{bmatrix} \nabla_{\beta old} L\left(\beta_{old}\right) \\ -\nabla_{\beta old} L\left(\beta_{old}\right) \end{bmatrix} + \begin{bmatrix} \lambda \\ \lambda \end{bmatrix}, \tag{5.62}$$

$$H\left(\beta^+,\beta^-\right) = \begin{bmatrix} H\left(\beta_{old}\right) & -H\left(\beta_{old}\right) \\ -H\left(\beta_{old}\right) & H\left(\beta_{old}\right) \end{bmatrix}. \tag{5.63}$$

The gradients and second derivatives inside the matrices are the same as defined above, i.e. $\nabla_{\beta old} L\left(\beta_{old}\right) = X'\left(Y - P\right)$ and $H\left(\beta_{old}\right) = -X'WX$. To start the iterative process for finding the optimal solution, we initialize the Active set such that it includes all variables on condition that they are greater than a threshold ε and the gradient at that point is less than zero,

$$\text{ActiveSet} = 1\left(\left(\beta^+,\beta^-\right) \geq \varepsilon, \nabla_{\beta^+,\beta^-} L\left(\beta^+,\beta^-\right) < 0\right). \tag{5.64}$$

The direction has then to be calculated adjusting the Newton formula,

$$\Delta_A = -H_A\left(\beta^+,\beta^-\right)^{-1} \nabla_A L\left(\beta^+,\beta^-\right), \tag{5.65}$$

while the step size λ is calculated iteratively using cubic interpolation. Afterwards,

$$\beta^A_{new} = \left[\beta^A_{old} + \lambda_{old}\Delta_{old}\right]^+. \tag{5.66}$$

The likelihood is updated and the procedure repeats itself until no further progress is observed such that the minimum is assumed to be found.[8]

5.3.5. ℓ_2 Regularized Logistic Regression

As already been mentioned above, when a linear function including an ℓ_2 type regularization constrain on the parameters is to be optimised, an analytical solution exists; thus the optimisation problem is straightforward to overcome. Under the Logistic loss function, though, there is no such convenient approximation to the problem of estimating parameters. The function which now becomes the focal point is

[8] See Schmidt *et al.* (2007).

$$\beta = \arg\min \left(\sum_{i==1}^{n} y_i \log p_i + (1 - y_i) \log (1 - p_i) \right) + \lambda \sum_{j=1}^{k} \beta_i^2. \qquad (5.67)$$

This is still differentiable, and a global optimum can easily be traced using a constrained optimisation algorithm where the constraint in this case is the regularization term, $\sum_{j=1}^{p} \beta_j^2 \leq t$.

5.3.6. Elastic Net Regularized Logistic Regression

The loss function belonging to this category incorporates both an ℓ_2 and an ℓ_1 regularization constrains, and for that reason no gradient can be calculated. To solve such a problem the method that has been adopted for the ℓ_1 regularized logistic regression will be used here as well modified however to incorporate the ℓ_2 term. To be more precise, the general representation of likelihood is the same with the one used before, i.e.

$$L\left(\beta^+, \beta^-\right) = f\left(\beta^+ - \beta^-\right) + \lambda \sum_i \left(\beta_i^+ + \beta_i^-\right) \quad \forall \beta_i^+, \beta_i^- \geq 0, \qquad (5.68)$$

where the f function here represents not the simple logistic function but includes an ℓ_2 constraint as well:

$$f\left(\beta\right) = \sum_{i==1}^{n} y_i \log p_i + (1 - y_i) \log (1 - p_i) + \lambda \sum_{j=1}^{k} \beta_i^2. \qquad (5.69)$$

The rest of the process towards the global optimum solution remains the same.

5.3.7. SPCA and the Logistic Function

When trying to perform a *Sparse Principal Components Analysis* while the logistic function is used the decomposition of the X matrix into $X\beta_{SPCA}$ is the same with that described in the previous section. The steps followed after the X_{SPCA} is obtained, though, deviate from the linear approach. Instead of regressing X_{SPCA} with Y to take the β_{OLS} estimates, we plug

the X_{SPCA} into the logistic probability function and calculate the relevant probability estimate for $\hat{Y}$.

5.3.8. Model Averaging using Variable Selection Algorithms

Alongside the algorithms presented above that deal with the problem of variable selection and the abundance of literature spawned as the aftermath of their proposition comes the intuitive question of which one of them performs better. In other words, beneath the problem of detecting the best predictors to achieve forecasting accuracy lies a more important consideration to be addressed first; that is model uncertainty. Model uncertainty has various implications in financial applications, and that is why more and more attention has been drawn to the resolution of the problem. Model averaging has still many to offer towards a safer approach for model selection also. Here two distinct approaches of averaging are investigated, Bayesian Approximation and Thick Modelling.

Bayesian Approximation, which has its traces in the Bayesian Model Averaging, was opted here once more. This is because BA overcomes complications embedded in the estimation process.

As discussed briefly in the introduction, when attempting to use algorithmic techniques to achieve asset predictability there exists uncertainty in the selection not only of the best model but also of the cutting point on the paths that this produces. In other words, every algorithm produces a collection of paths. These paths represent every new predictor that enters the forecasting equation, creating this way a new, less sparse model from the previous one. In order to select which of these sparse models generated at each step should be picked, we use model selection criteria, cross-validation etc. (for the case of LARS only we can use the C_p as an additional criterion). This procedure, however, introduces an additional 'cutting point' uncertainty which we address here, also using model averaging.

To summarize, model averaging in this chapter serves a double purpose. First, it is used to eliminate 'cutting point' uncertainty and second to account for model uncertainty, when deciding between competing algorithms.

5.4. Methodology

After laying the mathematical and conceptual foundations for some of the most popular algorithms used in variable selection problems, this section offers a more detailed analysis of the steps followed to solve our problem.

To start with, let us consider an unbiased investor that stands a time t and has available a wide number of predictors all including publicly available information and historical observations from time 1 to time t (where time 1 is considered the starting point of the history relevant to the investor's problem). If this investor wants to make a safe prediction for an asset at time $t+1$ by conditioning on all the information that he has gathered until time t, he faces an apparent dilemma. This is the fact that some of the candidate predictors are indeed important, while others are of diminishing economic significance. Traditionally, solutions to such a problem have one common direction. This consists of first finding as many as 2^p-1 combinations/models, where p is the number of available predictors, and based on these combinations regressing as many as 2^p-1 times (one for every candidate model). Finally, the method that in-sample performs better than its rival alternatives is be selected using various criteria. The criteria that rank the candidate models in-sample can be for example the R^2, MSE, Information criteria or any other financial criterion believed to be helpful when taking this decision (i.e. Sharpe Ratio etc.). This best performer in-sample is afterwards used for the $t+1$ out-of-sample forecast. The approach is the *all-subsets* method, and is considered by many the safest choice in comparison to the variable selection algorithms. But this method comes at a cost, as it dictates the number of regressions to be performed to grow exponentially as the number of available predictors increases, creating a sometimes unbearable burden for the researcher. Additionally, using the *all-subsets* technique the assumption that the best in-sample model will perform optimally out-of-sample also is unavoidably often incorrect. In conclusions the set up of the problem is as follows.

Let $Y = (y_1, y_2, \ldots, y_n)'$ be a vector including excess stock returns and X be a matrix of n observations and p_{M_i} predictors i.e. $X = (x_1, x_2, \ldots, x_n)'$, $x_{M_i} = \left(1, x_{j2_{M_i}}, \ldots, x_{jp_{M_i}}\right)$, where $j = 1, 2, \ldots, n$ and p_{M_i} varies depending on which model among the 2^p -1 used every time with p representing the total of all predictors. Using the OLS framework, the target becomes min-

imizing the Residual Sum of Squares using data from time 1 to time t for the dependent variable and from time 0 to time $t-1$ for the independent predictors, i.e.

$$\begin{aligned} \hat{\beta}_{M_x} &= \arg\min \sum_{i=1}^{n} \left(y_i^t - \sum_{j=1}^{p_{Mx}} \beta_{j_{Mx}} x_{ij_{Mx}}^{t-1} \right) \\ &\forall \quad M_x \quad \text{where} \;\; x = 1, 2, \ldots, 2^p - 1. \end{aligned}$$

After calibrating the betas and estimating the log-likelihood, the following criteria are used to rank the competing models:

$$\begin{aligned} R_{M_x}^2 &= 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_{it} - y_{it})^2}{\frac{1}{n}\sum_{i=1}^{T} (\hat{y}_{it} - \bar{y}_{it})^2} \qquad (5.70) \\ &\forall \quad M_x \quad \text{where} \;\; x = 1, 2, \ldots, 2^p - 1, \\ AIC_{M_x} &= \exp\left(\frac{2p_{M_x}}{n}\right) \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{it} - y_{it})^2, \qquad (5.71) \\ BIC_{M_x} &= n^{\frac{p_{M_x}}{n}} \frac{1}{n} \sum_{i=1}^{n} (y_{it} - y_{it})^2. \qquad (5.72) \end{aligned}$$

The best model depending on the ranking criterion every time is used to make tomorrow's forecast. This is to say,

$$y_{t+1} = \hat{\beta}_{M_{BEST}} X_{t_{M_{BEST}}}. \qquad (5.73)$$

This procedure is repeated in order to acquire $y_{t+2}, y_{t+3}, \ldots, y_{t+h}$, where h here represents the forecasting horizon, using a rolling window. For example, to make a prediction for y_{t+2}, we first drop one observation from the start and include the last observation we have used plus one new. Having updated our data, we are now ready to take the beta estimates using the following formula:

$$\hat{\beta}_{M_x}^t = \arg\min \sum_{i=1}^{n} \left(y_i^{t+1} - \sum_{j=1}^{p_{M_x}} \beta_{jM_x} x_{ij}^t \right)^2 \quad \forall M_x, \qquad (5.74)$$

where as before $x = 1, 2, \ldots, 2^p - 1$. To select the best model one of the

predefined criteria and is used. Finally, using the best model's estimated betas, the $t+2$ out-of-sample forecast can be made:

$$y_{t+2} = \hat{\beta}_{M_{BEST}} X_{t+1 M_{BEST}}. \tag{5.75}$$

The same procedure is followed for every step of the remaining targeted forecasting horizon also.

As an alternative to *all-subsets* we find *Ridge regression*. Although ridge regression can be analytically tractable, it does not produce sparse models. The method though can undoubtedly be regarded as a benchmark, and for that reason it will be used for comparison reasons. Again using data available for the independent predictors, we calculate the beta estimates and use them to predict the value of the dependent variable at time t+1 repeating the procedure using a rolling window until *t+h*. For the $t+1$ forecast the relevant betas are derived by minimizing the equation below:

$$\hat{\beta}_{ridge} = \arg\min \sum_{i=1}^{n} \left(y_i^t - \sum_{j=1}^{p} \beta_{ridge} x_{ij}^{t-1} \right)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2{}_{ridge}. \tag{5.76}$$

The equation has the tuning parameter λ that plays a defining role, since it not only prevents the betas from explosion but also can actually drive their values towards a particular direction. Taking into account all the above, cross-validation as a widely acknowledged practice is used to trace the optimal λ.

The alternative proposal to the investor comes from the algorithmic model selection family, the LASSO well discussed before. The inputs to the algorithm are the vector of the dependent variable $Y = (y_1, y_2, \ldots, y_n)'$ and the matrix of the independent predictors $X = (x_1, x_2, \ldots, x_n)'$ which includes all candidate alternative series to the researcher relevant to making that particular prediction of the dependent variable.

Again a rolling window is used to make predictions from time $t+1$ to time $t+h$.

For example, for time $t+1$ the target is minimizing the following penalized residual sum of squares:

$$\hat{\beta}^t_{M_x^{LASSO}} = \arg\min \sum_{i=1}^{n} \left(y_i^t - \sum_{j=1}^{p_{M_x^{LASSO}}} \beta^t_{jM_x^{LASSO}} x_{ij}^{t-1} \right)^2 + \lambda \sum_{j=1}^{p_{M_x^{LASSO}}} \left| \beta^t_{jM_x^{LASSO}} \right|, \tag{5.77}$$

where M_x^{LASSO} are the steps on the path that the algorithms produces, while this time we have much quicker convergence.

Each of these steps represents the additional predictor which enters the model starting from a model with no predictors but just a constant to one where all covariates have entered the predictive equation. At this final step the beta estimates are the OLS values. But in between the constant model and that produced by regressing all available predictors on the dependent variable there is a plethora of sparse candidate alternatives produced by LASSO ready to be used for forecasting. To select the 'cut off' point on the path, though, we again have to rely on the same criteria used before for the *all-subsets* method, i.e. the R-square, Akaike and Bayesian Information Criterion. To these an additional proposal to identify the best sparse model will be added, 10-fold cross-validation. Some further details in steps regarding cross-validation follow.

Initially the training set that has been used until now as a whole is divided into 10 equal subsets. Let T stand for 'training set'; then we partition that as below:

$$T = T_1 + T_2 + \ldots + T_{10}$$

Each of these subsets will be used as the pseudo out-of-sample period or else the 'test set' meaning that every time one of the ten subsets will be excluded, leaving only the remaining nine subsets as the new 'trimmed' training set. For example,

$$T^{trim} = T_1 + T_2 + \ldots + T_9$$

or

$$T^{trim} = T_1 + T_3 + \ldots + T_{10} \text{ etc.}$$

The second step involves fitting LASSO to the data of the 'trimmed' training set. As the algorithm does not produce a single solution but a path consisting of multiple steps, the target at each step is to make predictions for every observation of the test set (the test set to excluded from the total as described previously). The error for every observation is afterwards reported and at the end the Mean Squared Error of the step for this test set is estimated. This is repeated for all different test sets so that we end up with ten MSEs (one for every test set) for every step. The last thing left to do is to estimate the mean of these MSEs and the step that has the lowest mean is the winning 'cut off' point. In other words, every step (λ value) has the following MSE based on the number of target forecasts in the test period:

$$\text{MS CVE}_k^{\lambda} = \frac{1}{T_k} \sum_{(y,x)\in T_k} \left(y - y^{\lambda}\right)^2 \ \forall k = 1, 2, \ldots, 10, \tag{5.78}$$

where MS CVE_k stands for Mean-Squared Cross-Validation Error of the k *fold* and T_k is the number of observations of the k *fold*. The average error of the step is

$$CVE^{\lambda} = \frac{1}{10} \sum_{k=1}^{10} \text{MS CVE}_k^{\lambda}. \tag{5.79}$$

Finally we fit the model with the optimal λ that produced the lowest average cross-validation error to the entire training set and make the ultimate target forecast for $t + 1$.

Apart from LASSO, LARS is also used in our quest to accurately identify those predictors that can potentially produce powerful forecasting models and consequently achieve the forecasting optimality so much desired. The general set-up resembles that of LASSO, since the input matrix and the target vector are the same while the equation of interest still imposes the same restriction onto the parameter values (ℓ_1 norm restriction). LARS, though, moves a step forward, as not only can it be easily modified to produce LASSO solutions but also it can improve *forward stagewise regression*, with which it has a great resemblance. This is because instead of performing small steps (changes in λ value) when trying to identify the next most correlated to the residual vector variable (practice commonly used in *forward stagewise regression*), the algorithm moves in a direction which is equian-

gular with those predictors already incorporated in the model using larger steps (smaller though than *forward selection).* The above procedure is repeated until all covariates have entered the model. This in turn implies that (as LASSO and *stagewise regression* do) instead of a simple solution, that could indicate to the researcher the best model to be used out-of-sample to make forecasts for the depended variable, it instead produces paths. To select one single or else the optimal cut-off point on those paths, the same criteria as before have to be used, i.e. R^2 , the two *Information Criteria* and the *10-fold cross-validation.* After the optimal proposition is identified and the beta estimates have been acquired the prediction for time $t+1$ is ready to be made.

To the above algorithmic approaches one more will be added in the analysis which combines both ℓ_1 and ℓ_2 norm restrictions imposed on the parameters. This is the previously discussed LARSEN and the equation of interest is the following:

$$\begin{aligned}\beta^t_{M_x^{LARSEN}} &= \arg\min \sum_{i=1}^{n}\left(y_i^t - \sum_{j=1}^{p_{M_x^{LARSEN}}} \beta^t_{jM_x^{LARSEN}} x_{ij}^{t-1}\right)^2 + \dots \\ &+ \lambda_1 \sum_{j=1}^{p_{M_x^{LARSEN}}} \left|\beta^t_{jM_x^{LARSEN}}\right| + \lambda_2 \sum_{j=1}^{p_{M_x^{LARSEN}}} \beta^{2t}_{jM_x^{LARSEN}}.\end{aligned} \tag{5.80}$$

LARSEN is the last algorithm that is used in this analysis that has as output a collection of steps, each of them represented with the acronym M_x^{LARSEN} for different values of the tuning parameters. Unavoidably the optimal cutting point is established using the predetermined criteria (R^2, AIC, BIC, 10-fold cross-validation).

The last method to be proposed here is *Sparse Principal Components Analysis*, which indeed produces those factors (sparse components) from the initial matrix of predictors that can be used later to produce just one single model. The number of *Principal Components* to be used depends on the nature of the problem, though rarely more than six to seven factors are introduced. Here six predictors were considered enough to capture the relationships of interest. Finally, this sparse model is regressed on the

dependent variable to make future predictions.

Apart from trying to forecast the value of the dependent variable, our investor is interested in betting on the direction of the index. In other words, by creating an investment strategy and taking the right positions we want to investigate whether we can actually generate excess returns and ultimately profit. For this reason, apart from the linear regression framework we deviate and use the logistic regression instead.

5.4.1. Accounting for a twofold uncertainty: *Cutting Point* on the Path and *Model* Uncertainty

To perform averaging using one of the two predetermined techniques (Bayesian Model Approximation and Thick Modelling), we have to consider two different aspects of uncertainty generated by variable-selection algorithms. First, we find an endogenous uncertainty created by every model as a single entity, while at the same time we have the unavoidable uncertainty engendered by the surfeit of all those different variable-selection proposals. As a result, we have to approach averaging from two different angles. To account for the internal uncertainty we initially average across the steps of the paths of a single algorithm. But we also average across the steps of all paths generated from all variable selection proposals so as to address model uncertainty.

Averaging across the steps of the paths for an algorithm has the following intuition. All the way through the analysis it has been repetitively highlighted that these variable selection techniques themselves (except from SPCA) do not produce a single answer as to which is the most favourable model to use. Instead, they create paths, so that the researcher has to rely on other criteria to make the decision of the optimal cutting point. This in itself creates an intuitive 'step' uncertainty (the algorithm converges in steps).

To address that problem we first use Thick Modelling. At the outset, for every step of the variable selection algorithm we calculate the in-sample MSE, R^2, AIC and BIC and finally the Cross-Validation MSE. Depending on which of the above criteria we want to use, each time we rank the step performances and take a percentage of the best performing models to use out-of-sample to make the desired forecast. These percentages that determine the number of models-steps to be averaged vary between 10%, 20%

and 30%. Larger retention percentages could be adopted, but it is acknowledged that smaller fractions sufficiently outperform.[9] The rest of the step performances that are not at the top 10%, 20% or 30% are discarded and beta estimates for those steps will not be used. On the other hand, however, equal weights will be applied to the remaining top model and this averaged scheme that is generated at the end will be the proposal to be used out-of-sample to make the forecast for $t+1$ until $t+h$ using a rolling window.

The second proposal comes from Bayesian Model Approximation (BMA). To execute this type of averaging, we use the BIC information criterion as calculated above for every step of the variable selection algorithm under consideration. BMA uses all steps, not just the top performing, and to them weights -probabilities- are assigned using the following probability index.

$$\text{Prob}\left(M_{step}\,|X\right) = \frac{\hat{m}_{M_{step}}}{\sum_{i=1}^{n} \hat{m}_{M_{step}}} \quad \forall step = 1, 2, \ldots, n, \tag{5.81}$$

where M_{step} is the model produced if the cutting point on the curve was that step, $\hat{m}_{M_{step}}$ is the marginal as approximated by the exponential of the Bayesian Information Criterion and n is the number of steps the algorithm needs to converge (at that stage all predictors have been included).

Turning to the second uncertainty, we have the obstacle of selecting a variable-selection algorithm among many other candidates. For this reason we utilize averaging to push the barriers beyond model indecisiveness. In other words, we try to prove that with averaging, no matter how little we know about specific cases where particular variable selection techniques work better, and whether these claims are substantiated or not by different types of datasets under different circumstances, we can always achieve outperformance out-of-sample.

In order to use Thick Modelling, we still estimate all the predetermined model selection criteria (MSE, R^2, AIC, BIC, cross-validation MSE) for every step of every variable-selection algorithm (LASSO,LARS, LARSEN, SPCA, Ridge Regression), and then we combine the results in a cumulative table. It is useful to repeat that SPCA and Ridge Regression do not produce a path but give us a single model proposal. But they are still included in the

[9] Pesaran & Zaffaroni (2006).

total rankings, which at this stage is performed across all model alternatives. The top 10% or 20% or 30% performances according to different criteria every time are kept and equal probabilities are assigned to them so as to produce a new averaged model to be used out-of-sample.

Relatively similar procedures are followed for the BMA. The BIC Information Criterion is estimated for all steps and all models, as in the previous method, which is used to generate the cumulative table including all approximated marginal distributions. If every single marginal in that cumulative table is divided by the sum of all marginals, probability weights are created and can be assigned to every step-model alternative. It might be postulated that since BMA includes all steps generated from every algorithm to create the new averaged model this could easily create an overfitted scheme where everything is put together. However, probabilities related to poor performers are very close to zero (the fraction marginal to sum of marginals approaches to zero as the numerator decreases). The relevant equation to produce the weight of interest to be assigned is as below, with n however extended here to include the total number of all the steps of all variable selection algorithms generated as they move towards convergence plus 2 (one for SPCA and one for Ridge Regression).

$$\text{Prob}\left(M_i \left| \mathbf{X} \right.\right) = \frac{m_{M_i}}{\sum_{i=1}^{n} m_{M_i}} \ \forall i = 1, 2, .., n. \tag{5.82}$$

5.5. Data

Before presenting the data used for the purpose of this analysis, further consideration should be given to existing evidence that exchange rates can have robust predictability power over commodity prices (Chen, 2004, Chen *et al.*, 2008). Additionally, linkages between exchange rates and equities (Stavarek, 2004) also exist in the literature, but until now only weak dubious conclusions can be drawn. Engle *et al.* (2005, 2007) insist that modelling exchange rates as a random walk can be justified. This chapter tries to investigate how common macroeconomic predictors, along with other publicly available indexes and indicators used to forecast both exchange rates and commodity prices, can be utilized to predict not only these two obviously interrelated market products but also equities and bonds. Until now bonds were only

associated with stock prices (Lo and MacKinlay (1997), Baker and Wurgler (2009)). Never before has evidence suggested that all four asset categories can in fact be forecasted using a common basket of predictors and with the same success.

All predictors are selected after a thorough literature review. Data can be easily acquired from most data providers (Bloomberg, Datastream, Db Epsilon etc.).[10] They cover the period between January 1988 and December 2008 inclusive. In particular the macroeconomic indicators, can be downloaded from relevant governmental (national statistics) web sites. The frequency considered throughout the analysis is monthly. The total is 252 observations, divided into the training and the validation set. Since our target is forecasting as far as $t + 120$, the training period includes 132 data points while the remaining 120 observations are used for validation.

The target is first to predict US excess stock returns and for that purpose the S&P500 Index is used as an indicator, while the risk-free rate is represented by J.P. Morgan's 3-month US total return Cash index.[11] For the commodities the S&P GSCI index is selected instead (formerly known as the Goldman Sachs Commodity Index). The index is widely tradable and versatile, including a variety of raw or primary products and covering all commodity sectors. Additionally, the targeted FX index was the US Dollar Index (USDX) which actually represents the relative value of the US dollar against a collection of foreign currencies including the euro, Japanese yen, Pound sterling, Canadian dollar, Swedish krona and finally Swiss franc. Finally, for bonds the CSFB, World High Yield Index is selected to be forecasted.

All these four indices are used interchangeably as the target variable to be predicted and also as a stand alone predictor for the remaining three. The remainder of predictors, all of which in total constitute a collection of 19 macro and financial indicators, are as follows. Initially the macroeconomic predictors include first the index of Narrow Money Supply (M1 for the US and M0 for the UK) and second the Inflation Index as embodied by the EU Harmonized Index of Consumer Prices (HICP All Items). Narrow Money comprises of all physical money and currency along with other liquid

[10] For a more detailed description of the data see Table B.1. in the appendix.

[11] This index is used to track total returns (capital gains plus dividend) for constant maturity euro-currency deposits.

assets that the central bank possesses and represents the most restrictive (narrowest) definition of money, this of a medium of exchange.

On the other hand, there are the financial predictors including first the Term spread as represented by the yield of a 10-year US Treasury Bill minus the Federal Funds Effective Rate, the Risk Spread Index as approximated by a 3-month US Certificate of Deposit (FRCDS3M Index) minus a 3-month US Treasury Bill (FRTBS3M Index). Additional financial indicators include the Interest Rate Trend, defined as the monthly change of the average weekly yield on long-term government bonds, the January Dummy used to capture the January effect and finally the CVIX volatility index (CBOEVIX).

But the majority of predictors consist of interrelated commodity prices and exchange rates, which belong to five small commodity exporters. This idea that "commodity currency" exchange rates could be highly effective in predicting future commodity prices is introduced by Chen *et al.* (2008). However, this chapter moves a step forward by extending the idea to equities and bonds. It sets out to prove that there are common factors that influence global highly tradable indexes, and correctly identifying them could have a tremendous impact on our predictive capability. The reason for selecting smaller economies, though with sufficient exporting history and market-based exporting rates, is that these countries are price takers due to their small sizes. They are incapable of imposing price trends and as such even a small fluctuation in the commodity price has a significant impact on their exchange rates. In particular the countries selected are Australia, Canada, South Africa, New Zealand and Chile, so their exchange rates against dollar are used as predictors. For each of these countries, commodities represent a high percentage of their total export earnings, while relevant to these countries exporting products include Corn, Soy Beans, Silver, Gold, Gas Oil and Crude Oil.[12]

[12]For further details on the justification of those predictors please refer to the original paper: Chen, Rogoff & Rossi (2008).

5.6. Results

5.6.1. Linear Results

In the linear-regression case the focus is on forecasting excess stock returns. The rolling window used extends up to $t+120$, while at the end the average performance of every model or averaged scheme for all out-of-sample target forecasts is reported and conclusion are drawn.

For the purpose of the analysis the *all-subsets* method is considered as the benchmark. In theory an investor who could regress as many as 2^n-1 times (which represents the number of possible combinations of available predictors) and by using one of the widely acknowledged model selection criteria (MSE, R^2, AIC, BIC etc.) must be able to make a trustworthy prediction for tomorrow. For the 19 predictors that are considered here this number exceeds hundreds of thousands of regressions and if one more predictor is added to the model the number of required regressions could explode to over a million. Those regressions are actually executed though, to serve the purpose of the research, and also it is investigated whether the performance of *all-subsets* can be improved by trying to eliminate the uncertainty created when considering which model selection criterion is optimal by the Thick Modelling Averaging technique.

When the target forecast variable is **US excess returns**, the final ranking reveals the following (Table 5.1). First, the optimal cutting point for variable selection algorithms proves to be of immense importance. By changing the selection criterion which varies between MSE, R^2, AIC, BIC (which all share common ground) and cross-validation, great differences in performances are revealed. Cross-validation for all variable selection methods outperforms the remaining model selection criteria. On the other hand, *SPCA* and *Ridge regression*, which do not base their performance on such indicators, seem to underperform their rivals. Turning to the all-subsets method, the use of AIC or BIC criterion combined with *Thick Modelling* as an averaging technique offers the method an improved performance. If averaging is not imposed, however, this performance which is already not great, could be in jeopardy as bottom rankings reveal. Actually, by using any of the AIC, BIC, MSE, R^2 criteria, which are very popular among researchers, the investor could end up producing with certainty something very close to the worst error of all models that are considered in this analysis. Having

mentioned all the above, it is more than evident that the profits could be totally wiped out under any model selection criterion if *all-subsets* without averaging is chosen. The problem, though, of which alternative proposal should we opt from the plethora of all model selection algorithms that exist in the literature still remains. But more importantly, even if we make this selection we could hardly ever assert that this will be the right one out-of-sample for all various asset classes and time periods.

Averaging does offer a solution to the problem, since results indicate a catholic superiority of Thick Model Averaging using cross-validation for all three retention percentages (10%, 20%, 30% of best models) for all variable selection algorithms (LARS, LASSO, LARSEN). The interpretation of the result indicates that no matter what retention percentage is selected (restricted though into logical barriers, since there is hardly any meaning using large retention percentages) and no matter which variable selection technique the investor decides to trust, similar prediction error will be produced under the TM spectrum.

Additionally, regarding the uncertainty ("cutting point" on the path or model uncertainty) that is more probable to drive our prediction outcome towards the wrong direction, we notice the following. The right "cutting point" on an algorithm seems to be of greater importance when compared to the general model indecisiveness. That is, variable selection algorithms do produce reliable results if only we could identify the time when we have to interrupt convergence. In other words, when we average the steps of a variable selection algorithm it seems to yield better results no matter which algorithmic technique we have selected than when averaging the total steps of all methods.

Regarding which averaging alternative is better, we also notice that Thick Modelling outperforms Bayesian Approximation[13]. It is noticeable also that Thick Modelling using cross-validation to rank candidate schemes outperforms the other ranking criteria (MSE, R^2, AIC, BIC), either we when consider "cutting point" on the path uncertainty or general model uncertainty.

Averaging also improves the performance of *all-subsets* RMSE, but on the whole it is not a competent alternative. *Elastic Net* also shows an erratic behaviour, since the method is placed among the bottom performers. On

[13]A result supported also from relevant literature: Pesaran & Zaffaroni (2006).

the contrary, when used under the averaging spectrum and in particular under Thick Modelling with cross-validation, its performance stabilises at first places. But overall, *Elastic Net* using *Thick Modelling* with cross-validation outperforms all other alternatives. Notable is the fact, however, that the differences in the RMSE reported from all algorithmic variable selection methods is small after performing averaging, since without that only inconsistent conclusions can be derived.

The use of other asset classes, however, will greatly facilitate reaching a more catholic conclusion regarding model selection alternatives, ranking criteria and whether averaging is the right direction to follow in order to account for all the predetermined uncertainties.

The use of cross-validation for *LARS, LASSO, LARSEN* still produces the desired results for the **FX** and **Bond indexes**. On the other hand, these methods without averaging have a volatile performance on the total rankings table and averaging schemes should be opted instead. Cross-validation also demonstrates a constant outperformance in comparison with other model selection criteria when *averaging paths across all algorithms* is under scrutiny for these two targeted indexes. Again, however, only under the Thick Modelling spectrum *averaging paths across all algorithms* exhibits a more competent performance. Turning to comparing the averaged schemes, it is quite conspicuous that all top-ranking positions belong to averaging alternatives using Thick Modelling, while all retention percentages as long as they remain low constitute an optimal choice. Ridge regression still retains an average performance overall, while *SPCA* is ranked near the bottom.

To conclude, for the linear regression model analysis for the Equity, Bond and FX indexes, it is evident that the *all-subsets* method has two internal flaws. Firstly it imposes enormous computational burden and in real time can not possibly be considered as a way to facilitate investment decisions. Secondly, it is also highly unreliable. It depends greatly on the model selection criterion the researcher uses to determine the performance in-sample of all possible predictor combinations. This criterion does not guarantee that the selected optimal combination of covariates will also outperform out-of-sample. The results prove the point by placing the method in bottom position on the RMSE total rankings table. On the other hand, averaging offers the so needed performance stability of the variable selection

algorithms. However, *Thick Modelling* using cross-validation offers better results for any model under consideration. Differences in the performance of the method under other ranking criteria (MSE, R^2, AIC, BIC) are small for all model selection algorithms.

One the other hand, the Commodity index seems a real challenge to forecast. All in all, *Elastic Net* is still the optimal model selection technique after again the Thick Modelling Averaging method is applied. However, instead of cross-validation here, using any of the MSE, R^2, AIC, BIC criteria for various retention percentages proves to be a safer option. The same holds for the *all-subsets* using *Thick Modelling* and the previous mentioned model selection criteria. This way from the bottom of the ranking the *all-subsets with averaging* methods regains a competent performance. Additionally, while unsurprisingly the Ridge Regression method is still an average alternative, *SPCA* proves to exhibit reinforced predictive power while LARS and LASSO move some places down in the top ranks.

To validate the predictive power of the models used we also performed the Diebold-Mariano test. This is in order to answer the question whether the predictions of a given model, A, are significantly more accurate, in terms of a loss function $g(\cdot)$, than those of the competing model, B. The loss function here in particular is the MSE, at 95% confidence level. Comparative results for all asset classes are reported in Tables 5.21-5.24. For all four different asset categories the all-subsets method using R^2 or MSE, with and without model averaging, generates significantly more accurate forecasts against the other alternative model predictions. Also models that completely failed to pass the test (their predictions were never statistically better than their rivals) were not included in the results tables.

5.6.2. Logit Results

On the other hand, when the logit regression is under scrutiny we are not interested in forecasting the actual asset stock return but instead we want to predict just the direction of the movement of the index at time $t+1$. This can also spawn simple trading strategies. By identifying whether the asset will rise or fall the following day, the investor can take the right position and earn profits no matter what will happen to its value the next trading day. Making the right choice, though, involves high earnings (or at least the potential to

make such when circumstances are favourable – lower transaction costs, no trading barriers etc.). But taking the wrong position can be devastating for the investor. This is why logit regression plays such an important role and an even higher responsibility is placed on averaging techniques to diminish the probability of a spurious result due to model or 'cutting point' uncertainty.

Some general conclusions with regards all four individual asset classes do exist for some variable selection techniques while for others performances vary depending on the nature of the targeted index (Table 5.2). In general the *all-subsets method* does not outperform any model selection algorithm or averaging scheme. This is for all four different asset categories. Especially, the lowest probability of successful prediction of the direction of the four indexes is demonstrated by the *all-subsets method without averaging*, while when averaging is executed the technique seems to produce improved results. It is also notable that the overall probability of successful prediction of the direction on the index each time using averaging with all-subsets tends to coincide for various retention percentages for the Thick modelling averaging method and for different model selection criteria (MSE, R^2, AIC, BIC). This fact clearly indicates a constant performance of the averaging scheme that is not dependent on the retention percentage or on the model selection criterion. On top of that, it improves the overall performance of the *all-subsets* method.

Additionally, at the bottom of all four asset classes we find a common technique, *averaging across all algorithms*. This conclusion is so strongly supported by empirical results that we could safely conclude that the method should be treated with consideration. On the contrary, *LASSO* is a top performer for all targeted indexes with percentages of successful prediction of the direction being improved when averaging is performed. *LASSO* appears more competent especially when performing *Thick Modelling with cross-validation*, while various retention percentages generate similar if not identical results. Moreover, *Bayesian Approximation* for all but the FX index is not a competent averaging alternative, with *Thick Modelling* demonstrating a more consistent performance overall. *Elastic Net* also demonstrates a superior record of successful forecasts when *Thick Modelling* is applied to the method. However, irrespective of the model selection criterion, if averaging is not applied the performance of the algorithm diminishes. Also results show that for all retention percentages *Elastic Net* under the *Thick*

Modelling spectrum produces identical results, eliminating this unavoidable internal uncertainty created from the technique (i.e. correctly selecting the optimal retention percentage).

Moving to the particular asset classes, it is notable that the **FX** and **Bond indexes** demonstrate a similar ranking table, while the overall percentage of their top performer method is higher than the equivalent top performer for the Equity or Commodity Index. The argument for this result can be either that the Equity and Commodity indexes are harder to predict, or the set of predictors we used for all indexes is more related to predicting FX and bonds. But the latter though cannot actually be substantiated, for the following reason. The majority of these predictors used for the purpose of our research include FX indicators as suggested by relevant literature which connects commodity prices with foreign exchanges.[14] In conclusions, the commodity index is the hardest to predict among the remaining asset classes.

Returning to Bonds and FX logit results, we can observe that *LASSO* with averaging is the best predictive method, while *Elastic Net* is also a top performer for these two asset classes. For the Bond index, though, *Elastic Net* with *Thick Modelling* using *AIC* and *BIC* as model selection criteria works better. On the contrary, for the FX index *Thick Modelling* with *MSE* and R^2 is preferable. On top of all, Bayesian Approximation under the LASSO spectrum for FX is the definitive method. At the bottom of the rankings we find, apart from the method of *averaging across all algorithms*, the *all-subsets method* for both asset categories. On the other hand, *Ridge regression* demonstrates a consistent average performance occupying places in the middle of the rankings.

For the **Equity and Commodity indexes**, we find that LASSO with cross-validation as well as *LASSO with Thick Modelling* using *AIC* and *BIC* as model selection criteria is preferable. For the Equity index, the *Ridge regression* scheme is still an average performer, but for Commodities it outperforms all other alternatives if combined *with Thick Modelling Averaging. Elastic Net* with *Thick Modelling* using *AIC* and *BIC* is still the second-best alternative for both targeted indexes. At the bottom, for Commodities we find as before *Thick Modelling across all algorithms* and the *all-subsets method.* However, for Equities and equities only the *all-subsets*

[14]Chen, Rogoff & Rossi (2008).

method demonstrates a better performance reaching top ranking positions, but only when is combined with the *Thick Modelling* averaging technique.

5.6.3. Investment Strategies: Mean-Variance Analysis

Modern Portfolio Theory can be naturally encapsulated in Markowitz's (1952) Mean-Variance approach, according to which the investor tries to identify the portfolio that offers the maximum expected return with the minimum possible variance. The logic behind this proposition, although usually condemned for its simplicity by advocates of behavioural finance, still remains a competing tool against more sophisticated approaches. This is not only due to its simplicity of implementation but also to its effectiveness on generating profits if combined with continuous rebalancing and accurate selection of investor's risk aversion. In 1990 Black and Litterman introduced a revolutionary (for at time) proposition to the investor, where he could actually incorporate to the objective pragmatic historic observations his subjective views for the future. In the Black-Litterman world, Markowitz's assumptions for normality of the market and linear dependent scenario views still hold. Normality assumptions not only of the past data under consideration but also of investor's beliefs render the validity of theory once more questionable and ultimately inadequate to capture all those irregularities that exist and define modern capital markets.

Extensions to the BL theory are rarely reported in the past decade, with the only recent exception the Entropy Polling theory as introduced by Attilio Meucci (2008b). In his ground-breaking approach, markets are described not only from returns but also from higher moments (implied correlation, volatilities etc.). In addition, there is a noticeable departure from normality and market equilibrium, since skewness, fat tails and co-dependence of extreme events are undeniably present. Finally, he introduces a significant improvement in the formulation of investors' views which are no longer restricted to be linear and are not only related to future expected returns. Sophisticated asset managers can and do have expectations about correlation, volatilities, tail behaviour etc. which can under this spectrum be formulated not only as equalities but inequalities as well. Entropy polling, although in its infancy, has huge prospects and a lot to offer to those that dare to involve themselves with pioneer investment management techniques. The only dis-

advantage is, however, the indisputable complexity of implementation. On the contrary, as has already been stressed, both Markowitz Mean-Variance portfolio theory and the Black-Litterman further development have prospered in the past as a result of their intuitive and uncomplicated nature. This fact, in conjunction with their competence in generating profitable results - even though their conceptual framework is founded under strict conceptual restrictions - make the two theories of undisputable significance for day-to-day asset management decisions. For all those reasons discussed above the fundamental mean-variance framework is employed and results of the implementation of this investment strategy is reported.

Having under our discretion 4 different asset classes, optimal allocation of the investor's wealth between them automatically becomes the focal point of this section, with primary objective the attainment of maximum profitability. The number of asset classes available that comprise a portfolio can in reality far exceed this number. However, it is strongly supported by empirical evidence that the fewer classes the more probable is predictability patterns of those model selection algorithms used in our research to be disclosed. As the number of asset classes increases results can be distorted. Since we are using monthly data, the rebalancing of our portfolio was also performed on a monthly basis. Also for the assessment of the overall performance three additional statistics are used along with the cumulative wealth (assuming that at time 0 the investor holds 1 unit of wealth): the Fama French Alpha, the Carhart Alpha and finally the Sharpe Ratio. All these metrics are widely used in the literature for assessing the performance of the various asset allocation strategies, while for further details of each distinct measure refer to the original papers.[15]

In addition, regarding the short sales constrain two distinct cases had to be reviewed. First, we consider markets where regulatory agencies that instigate short sale restrictions do exist. Under those restrictions, the performance of our portfolio under the Mean-Variance spectrum was assessed, using all those four measures above. This scenario, although possible under extreme market conditions (see September/October 2008) or due to other market irregularities (liquidity constrains, non synchronous trading), can have a categorical effect on profitability and should only be chosen after careful consideration. For this reason the second more rational view that

[15]Fama & French (1992), Carhart (1997).

we investigate involves the exclusion of the no-short-sale restriction but imposed a second constraint, this time that the investor can either sell or buy at a maximum of 100% of the value of the particular asset class. Confining the percentage of leverage can still diminish the overall performance of the strategy, but it was considered more appropriate to pragmatically simulate an investment strategy that can mirror the real-world as accurate as possible. In a real-time environment it is often the case that taking a position for example 300% long or short on an asset is not possible. Again the strategy was assessed using the four performance measures for all available variable selection and averaging methods investigated in this chapter.

5.6.3.1. Results - short sales restriction

In terms of cumulative wealth (Table 5.3), although the *all-subsets* method using the Akaike or the Bayesian Information Criterion is ranked at the top, using the MSE or the R^2 criterion instead could dramatically reduce the overall portfolio's performance. Averaging with *Thick Modelling* under the same method does produce equivalent results while once more the retention percentage selected plays a trivial role in influencing the outcome of the process. A stronger evidence of the importance of averaging methods overall can be traced at the second-best alternative regarding cumulative wealth, which is *Elastic Net*. *Elastic Ne*t under the *AIC* criterion produces a competitive investment result, while this is not the case for the same method under the BIC criterion, since it is placed near the bottom of the total ranks. *Averaging with Thick Modelling* using *Elastic Net* on the other hand performs sufficiently well, and most importantly irrespective of the model selection criterion (AIC, BIC, MSE, R^2) or the retention percentage. All *Thick Modelling* alternatives produce similar cumulative wealth. *SPCA* also exhibits a noticeable performance, while *LASSO* and *LARS* produce mediocre results. Also *averaging across all algorithms* does not generate the desired outcome.

On the other hand, Sharpe Ratio indicates a similar ranking for all methods. *All-subsets* is still ranked at the top, but caution should be placed on correctly identifying the optimal model selection criterion which in turn necessitates the introduction of averaging. *LASSO* here has a more dominant presence, while *Elastic Net* loses some strength. But for the Sharpe Ratio

which is produced under the *Elastic Net* variable selection method, results do vary whenever the model selection criterion is altered. Averaging more is again necessary in order to secure stability.

Turning to rankings according to Fama French alpha and Carhart alpha, results are almost identical. This is due to the fact that the four-factor model of Carhart is a projection of the three-factor model of Fama and French (1993) plus an additional factor for the momentum as suggested by Jegadeesh and Titman (1993). A positive alpha is an indicator of a better than expected performance under the portfolio's systematic risk. So the higher this alpha above the zero threshold the better. In contrast with the previous two performance measures (cumulative wealth and Sharpe Ratio) here *LARS*, *LASSO* and *Ridge Regression* are top performers. *Elastic Net* also remains a competent scheme, but averaging methods here do not add anything to the overall adequacy of the method. The same for *all-subsets*: without averaging it is still among the first choices for producing portfolios with higher positive alphas. Also *SPCA* proves to be inefficient to fulfil the same purpose of successfully providing asset allocation solutions that earn higher than anticipated returns.

5.6.3.2. Results - no short sales restriction

By partly removing the short sales restriction and imposing instead a leverage cap at 100% both when positioning long or short on an asset the total rankings do not changed significantly (Table 5.4). However it is noticeable that cumulative wealth does increase considerably for all methods. This comes to substantiate the claim of what a significant impact on profitability such kind of restrictions can make. The best scheme (if the initial assumption is that our wealth equals one for example) produces as much as 4.31 cumulative wealth in a ten-year period (all calculations are performed purely out-of-sample). This is generated by all-subsets method with *Thick Modelling* as an averaging proposal for all retention percentages (these vary from 10% to 30%) using the *AIC* or the *BIC* model selection criterion. The same holds for the Sharpe Ratio, with the highest value for that measure being produced again from all-subsets with *Thick Modelling* and *AIC* or *BIC* proposition. However, *Elastic Net under the Bayesian Approximation* now exhibits a superior performance, while the method without averaging

is placed all over, from top to bottom ranking positions. This repetitive uncertainty only reconfirms the importance of averaging.

Regarding the Fama French and Carhart alphas, *Elastic Net*, *LARS* and *LASSO* constitute the top choices for producing portfolios with competent performance. But it is noticeable that averaging schemes for these two performance measures do not significantly improve the overall portfolio's alphas for all methods. They do, however, generate augmented alphas for some particular models, and their role should not be disregarded. *Ridge regression* exhibits a competent performance here for all four metrics, while *SPCA* is not a strong alternative.

5.6.4. Logit based Analysis

5.6.4.1. Results - short sales restriction

For ***Equities*** when short sales are not allowed (Table 5.5), *Thick Modelling* with *all-subsets* using the R^2 or *MSE* model selection criterion and *Thick Modelling* with *Elastic Net* under the R^2 or *MSE* or *cross-validation* is the best alternative in terms of cumulative wealth over a ten-year period. Once more the selected retention percentage plays trivial role since rankings are identical. However, although *Thick Modelling* with *all-subsets* using *BIC* or *AIC* has similarly profitable results, *Thick Modelling* with *Elastic Net* under the same criteria produces ambiguous conclusions. For some retention percentages the method generates average outcomes, while for others it performs poorly. There is a pattern, though, since without any averaging still *Elastic Net* under R^2, *MSE* and cross-validation is better than under the remaining two information criteria used for model selection. Also without averaging *Elastic Net* generates average cumulative wealth, which is considered by most the ultimate investor's objective. Also *LASSO* with cross-validation is ranked at a high position in terms of wealth generation, with the significance of the method being more apparent in the remaining three ranking criteria (Fama French and Carhart alpha and Sharpe Ratio). Additionally the performance of the algorithm is improved under the *Thick Modelling Averaging* approach when cumulative wealth is used as a measure of the efficacy of the technique.

In conclusions *Thick Modelling* offers enhanced cumulative wealth for the majority of the algorithms investigated. Averaging across all models

produces mediocre performance when cumulative wealth and Sharpe Ratio are used as indicators, while the alphas (both Carhart and Fama French) are significantly lower than the alphas produced by the other competing alternatives. Finally, for equities when short sales restriction is imposed, SPCA should not be opted since for all performance measures results are equally poor.

Comparing the above results with those of the section where the percentage of successful prediction of the direction of the index is discussed, we have to notice the following (see Table 5.13). The top performing models exhibit a consistency in their performance. Under this context the highest cumulative wealth is produced by models that also exhibited the highest rate of accurately identifying the direction of the asset move. These top performers are represented by *Thick Modelling* with *all-subsets* using the R^2 or *MSE* and *LASSO* with *cross-validation*. *Elastic Net* on the contrary demonstrates an irregular performance pattern, with some variations of the method [16] retaining their strength while others not.

Turning to **Bonds**, what is conspicuous is that the averaging schemes for all different algorithms prevail (see Table 5.6). *Thick Modelling* with *all-subsets* using the R^2 or *MSE* model selection criterion is again at the top in terms of cumulative wealth and Sharpe Ratio, together with *LASSO* using *Thick Modelling* in combination with the *AIC* or *BIC* instead. *Elastic Net* when combined with *Thick Model Averaging* under *AIC, BIC or cross-validation* is also among the optimal choices in terms of its ability to generate profits. Of the worst performers we identify the *Ridge regression* and *averaging across all algorithms*. It is evident also that the last scheme consistently underperforms other methods under both the linear and the non-linear spectrum across all asset classes. Ridge regression also produces mediocre results, with few exceptions. Additionally *LASSO* with *Bayesian Approximation* for bonds is in the top four positions when wealth is used as a performance indicator. *SPCA* on the other hand, as in equities fails to generate competitive results when profit and Sharpe Ratio are used as indicators. However, *SPCA* manages to yield portfolios with high alphas (both Fama French and Carhart alpha). Also when the performance measure examined is the portfolio's alphas *Elastic Net* with *Thick Model averaging*

[16] The term "variations" of a model selection algorithm implies if averaging or not is applied and which of all the different model selection criteria is chosen.

and *AIC or BIC* selection criterion is at the top of the ranking tables along with *LASSO* under the same averaging regime. *All-subset*, however, produces lower, though positive alphas, which in all cases is an indication of a performance above expectation. Once more, *all-subsets with Thick Modelling* is better than without averaging. *Ridge regression* as before generates middling alphas and even lower Sharpe Ratios under all different variations of the algorithm (with or without averaging, different model selection criteria, retention percentages for *Thick Modelling* etc). All in all, however, averaging alternatives to single model prevail and we additionally observe a consistency of performance for the *Elastic Net*, *LASSO* and *all-subsets algorithms*.

Turning our attention to results of the previous section when a logistic regression for prediction of the direction of the Bonds index is performed for a window of 120 months, we observe the following (see Table 5.14.). *LASSO* and *Elastic Net* using *Thick Modelling* with *AIC or BIC* are the optimal forecasting methods, which is also apparent when these algorithms are used to facilitate portfolio investment decisions. *Bayesian Approximation* under the *LASSO* continuum is also top performer when used to predict the direction of the index and also when employed for investment reasons).

When **Commodities** are under consideration while still the short sales restriction holds we observe the following (Table 5.7.). For all four portfolio performance measures the averaging suggestions are at the top of the rankings. Thick Model average when combined with *LASSO* as well as *all subset regression* yields the highest cumulative wealth. *LASSO with Thick Modelling* also has the highest portfolio alphas, while Sharpe Ratio for the method is a little lower but still positive and at a satisfactory level. On the other hand, although *Elastic Net* produces significantly high alphas the Sharpe Ratios and cumulative wealth for the various alternatives of the method are low.

An outcome worth noticing is that *averaging across all algorithms* here exhibits a significant improvement in terms of cumulative wealth and most importantly Sharpe Ratio. However, the alphas reported for the method are still near zero, indicating a poor management of the systematic and unsystematic risk. However, the fact that the alphas are even marginally different from zero, the cumulative wealth near 2 for a ten-year period and the Sharpe Ratios higher than the other methods means that the method

cannot be so easily disregarded. Its potential as a way to eliminate model uncertainty is also validated for a different asset class that will follow next in the analysis, the Foreign Exchange index.

On the contrary, with *averaging across algorithms, Ridge regression* here generates high Fama French and Carhart alphas, while in the ther two performance indicators the results are not exceptional but still acceptable. Additionally, although *SPCA* and *Bayesian Approximation* on average are ranked in the middle (with *Ridge regression and BA* being the most preferable choice), competing alternatives should be opted as more consistent performers instead.

In comparison with the results from the logistic regression (Table 5.15), it is evident that *Ridge regression* exhibited competent percentages of successful prediction of the direction of the index in the total out-of-sample period, while on average it could not retain this superiority when applied in an asset allocation framework. On the contrary, *LASSO* and *Elastic Net with Thick Modelling* consistently outperformed relevant alternative propositions.

Finally, the analysis of the investment strategies concludes with results related to the **Foreign Exchange index** (Table 5.8). *LASSO and all-subsets regression with Thick Model Averaging* produce the highest wealth in a ten-year investment horizon. Most importantly, however, is the fact that for the first time we observe that Fama's and Carhart's alphas take for all methods near zero to negative values while the same holds for Sharpe Ratios. This can signal a potential inability of the algorithmic techniques to capture adequately the dynamics of this particular asset class, or as others have claimed (for example Engle and West, 2004), this could imply a general randomness of the exchange index such that forecasting is difficult. However, cumulative wealth, although lower in total when compared to equivalent results from other asset classes, is still positive for the various models. *Bayesian Approximation and LASSO* demonstrates an improved performance, while *Elastic Net with Thick Modelling* remains a competent option in terms of net selectivity performance results (selectivity minus diversification as this is represented by portfolio's alphas). The same holds for *averaging across all algorithms*. The scheme produces marginally above-zeros alpha when others underperform. *Ridge regression* in accordance with related results to the stock and bond index generates low values for all four performance indicators, while *SPCA* is at the bottom of the rankings with negative alphas

and Sharpe Ratio and total wealth marginally above 1.

When returning to the rankings of the average percentage of correctly forecasting the direction of the index using logistic regression (Table 5.16), it is conspicuous that *LASSO with Bayesian Approximation* as well as *LASSO with Thick Model averaging* is at the top of the ranks, and also *Elastic Net*, under both averaging proposals. This stands as the empirical testimony of the superiority once more of the averaging schemes and the consistency in the performance of the two variable-selection methods. The same can be stated for *all-subsets with Thick Modelling*. But still, negative alphas and Sharpe Ratios as performance indicators cast a shadow on the results and the predictive competency of all proposals for investment problems related to foreign exchange.

5.6.4.2. Results - no short sales restriction

The next step in the analysis is to remove the short sales restriction. However, it was considered wiser to avoid taking any huge leverage and restrain the weights to $\pm 100\%$. Under this regime, we investigate how the rankings altered for the various asset classes we are considering.

First we look at the ***Equities index*** (relevant results are provided in Tables 5.9 and 5.17). Initially we observe a significant increase in the cumulative wealth and also in the Sharpe Ratios for all models as opposed to the ranking results when short sales were not allowed. However, the Fama French and Carhart alphas drop considerably in value, and for those schemes that underperformed, negative alphas are reported. Those models relate mainly to the method of a*veraging across all paths of all algorithms* as well as to *SPCA*.

Also it is worth mentioning that a grouping effect is observed. In other words, models related to the same algorithm but under different variations such as averaging with different model selection criteria or *Thick Modelling* with various retention percentages exhibit a common or very close performance. This brings us closer to the attainment of a solution to the model selection dilemma which could be related also to which averaging method should be employed, unless we chose a single model each time.

Having in mind all the above observational results, we notice in a more precise manner that *all-subsets with Thick Model Averaging* under all four

portfolio performance measures is the most favourable choice. *LASSO with cross-validation* and without any averaging comes next in terms of cumulative wealth and Sharpe Ratio. However, the method generates negative alphas, so this erratic behaviour in the results should make the investor more sceptical about the competence of the algorithm. On the other hand, *LASSO with Thick Modelling* for lower retention percentages (10%, 20%) generates more constant performance based on all four ranking criteria. Additionally, *Elastic Net with Thick Modelling* retains and enhances its predictive power in comparison with the previous section where short selling is forbidden. *Ridge regression*, on the other hand, has again an average effect.

Turning to the **Bond index**, related asset allocation results (Tables 5.10 and 5.18), we observe that once more *Thick Modelling with LASSO or Elastic Net* is at the top when cumulative wealth and Sharpe Ratios are used as performance measures. However, alphas under both the three-factor model of Fama and French and the four-factor alternative of Carhart are in the majority of the cases negative. So, again depending on the interpretation of these indicators, we can conclude that the performance of the modelling strategies did not live up to the anticipated returns. However, one cannot disregard the other two measures, especially when cumulative wealth related to the total investment horizon increased in comparison with results when short sales restrictions are imposed. Once more the "grouping" effect observed in equities (where the same algorithm under different variations of the averaging strategies generates identical results) is also evident for this asset class. Bottom performer remains *averaging across all algorithms*.

For the **Commodity index**, however, as opposed to bonds, the outcome of the asset allocation strategy is completely different (Tables 5.11 and 5.19). Here we observe an increase in the cumulative wealth of around 100%, as well as an improvement in the other three performance indicators. This result could also relate to the nature of the data, since the majority of predictors found in relevant literature are used for forecasting commodity prices. Alphas are back to positive for those top performing schemes which include *Thick Model Averaging* strategies under *LASSO, all-subsets regression* and *Elastic Net*. These methods consistently perform with or without short sales restriction, while the same holds for *SPCA*, which retains its ability to generate profitable investment strategies with performance higher than the market would dictate. Here, however, as opposed to the restric-

tive asset allocation case the *averaging across all paths of the algorithms* is ranked at the bottom. But this result is in accordance with the general incapability of the method to generate competent results, as has been reported for different asset classes before, both for the linear and the non-linear models. Finally, *Ridge regression* retains average predictability. Also important is the fact that *Bayesian Approximation and Thick Modelling* assists in the attainment of higher performance as opposed to equivalent alternatives without averaging.

In conclusions, the **Foreign Exchange index** under no short sales constraint generates, for all asset classes, identical rankings with those reported when short sales are not allowed (Table 5.12.). What is also apparent is the fact that alphas and the Sharpe Ratio for top-performing models are increased and non-zero. However, negatives do exist, so careful consideration should be given when selecting a method that relates to negative performance indicators. An example *all subset regression*; although for some variations of *Thick Modelling* the method can be regarded as a good choice, the alphas generated under all alternatives are negative, so the investor should be cautious under this variable selection spectrum. Cumulative wealth is also distinctively increased for the top models. Under this measure, and also under Sharpe Ratio, *Bayesian Approximation and Thick Model Averaging* primarily for *LASSO* and subsequently for *Elastic Net* are a successful determinant of the investment result. The same outcome is observed in the no-short-sales investment allocation case.

Last of all, detailed results that report the change in the ranking position of a method, first when its performance is based on the result of the percentage of successful prediction of the direction of the index[17] and afterwards on the asset allocation strategy (for all four performance measures independently), are provided in Table 5.20.

5.7. Conclusions

In this chapter we consider two of the most common problems encountered in econometric applications. The first refers to the variable selection problem, while the second relates to model uncertainty. In an attempt to

[17] Average performance for the total investment horizon as this was determined by the logistic regression.

address this obstacle we utilize some robust algorithmic techniques commonly used in scientific domains outside finance. The methods selected include the all-subsets, Ridge regression, LASSO, LARS, Elastic Net and finally Sparse Principal Component Analysis. On the other hand, model uncertainty is addressed by averaging methods, including Thick Model Averaging and Bayesian Approximation. The analysis is performed under both the linear and the non-linear regression framework. For the empirical validation a collection of nineteen predictors, all publicly available, were used while attempting to forecast the return of various asset classes (stock, bond, commodity and foreign exchange index), as well as the direction of their movement. Finally, asset allocation strategies are applied first with and afterwards without a short-sales restriction.

Results in all cases indicate the superiority of those algorithms where averaging methods were employed. In particular *LASSO* and *Elastic Net under Thick Model Averaging* perform better than rival alternatives, while on the other hand we find *averaging across paths of all models* consistently underperforms. *Ridge regression*, though not by definition a variable selection technique, and used to restrain any unnecessary parameter explosion, does not on average produce competent forecasting and investment-related results. Finally, *all-subsets* regression, although it demonstrates high potential, imposes an unwanted computational burden on the researcher, which can easily be avoided if other variable selection alternatives are chosen instead.

Rank	Equity	RMSE	Bonds	RMSE	Commodity	RMSE	FX	RMSE
1	en 20% cross	4.42%	lars cross	2.94%	en 30% aic	7.28%	lars 30% cross	1.14%
2	lasso 20% cross	4.42%	lasso cross	2.94%	en 20% aic	7.28%	en 30% cross	1.14%
3	20% across algorithms cross	4.42%	10% across algorithms cross	2.95%	en 10% aic	7.29%	lasso 30% cross	1.14%
4	lars 20% cross	4.42%	en 10% cross	2.95%	en 30% mse	7.29%	30% across algorithms cross	1.14%
5	10% across algorithms cross	4.42%	lasso 10% cross	2.95%	en 30% R^2	7.29%	lars 20% cross	1.14%
6	en 10% cross	4.42%	lars 10% cross	2.95%	en 20% mse	7.29%	en 20% cross	1.15%
7	lars 10% cross	4.42%	en 20% cross	2.95%	en 20% R^2	7.29%	lasso 20% cross	1.15%
8	lasso 10% cross	4.42%	lasso 20% cross	2.95%	en 10% R^2	7.29%	20% across algorithms cross	1.15%
9	en 10% bic	4.42%	20% across algorithms cross	2.95%	en 10% mse	7.29%	lars 10% cross	1.15%
10	en 30% cross	4.42%	lars 20% cross	2.95%	en 30% bic	7.29%	10% across algorithms cross	1.15%

Table 5.1.: Rankings of the first 10 best Linear Models and averaging schemes for all asset classes.

Rank	Equity	RMSE	Bonds	RMSE	Commodity	RMSE	FX	RMSE
1	lasso cross	63.3%	lasso 10% aic	64.2%	ℓ_2 10% aic	60.0%	lasso BA	65.0%
2	en cross	60.8%	lasso 10% bic	64.2%	ℓ_2 10% bic	60.0%	en 30% mse	62.5%
3	all subset best20% mse	60.0%	en 10% aic	62.5%	ℓ_2 20% aic	60.0%	en 30% R^2	62.5%
4	all subset best20% R^2	60.0%	en 10% bic	62.5%	ℓ_2 20% bic	60.0%	en BA	62.5%
5	all subset best30% mse	60.0%	en 20% aic	62.5%	ℓ_2 30% aic	60.0%	lasso 10% cross	62.5%
6	all subset best30% R^2	60.0%	en 20% bic	62.5%	ℓ_2 30% bic	60.0%	lasso 30% mse	62.5%
7	lasso 30% aic	60.0%	en 30% aic	62.5%	lasso cross	58.3%	lasso 30% R^2	62.5%
8	lasso 30% bic	60.0%	en 30% bic	62.5%	ℓ_2 BA	56.7%	lasso cross	62.5%
9	all subset best aic	59.2%	lasso 10% cross	62.5%	ℓ_2 10% cross	55.8%	en 10% cross	61.7%
10	all subset best bic	59.2%	lasso 30% aic	61.7%	ℓ_2 bic	55.8%	en 10% mse	61.7%

Table 5.2.: Rankings of the first 10 best Logit Models and averaging schemes for all asset classes.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	all subset best bic	1.70	lars bic	0.91	lars bic	0.88	all subset best bic	0.06
2	all subset best30% aic	1.58	lasso bic	0.91	lasso bic	0.88	all subset best30% aic	0.05
3	all subset best10% aic	1.56	en bic	0.87	en bic	0.85	all subset best10% aic	0.05
4	all subset best20% aic	1.55	en cross	0.85	en cross	0.84	all subset best20% aic	0.05
5	all subset best aic	1.47	lars cross	0.77	lars cross	0.76	all subset best aic	0.04
6	all subset best10% bic	1.47	lasso cross	0.77	lasso cross	0.76	all subset best10% bic	0.04
7	all subset best20% bic	1.42	ℓ_2	0.70	ℓ_2	0.70	all subset best20% bic	0.03
8	all subset best30% bic	1.40	en R^2	0.69	en R^2	0.68	all subset best30% bic	0.03
9	en aic	1.39	all subset best R^2	0.69	en mse	0.68	en aic	0.03
10	lars aic	1.29	all subset best mse	0.69	all subset best R^2	0.61	lars aic	0.02

Table 5.3.: Rankings of of the first 10 best Linear Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	all subset best10% bic	4.31	en bic	1.41	en bic	1.34	all subset best10% bic	0.17
2	all subset best20% bic	4.26	lars bic	1.35	en cross	1.31	all subset best20% bic	0.17
3	all subset best30% bic	4.19	lasso bic	1.35	lars bic	1.27	all subset best30% bic	0.16
4	all subset best10% aic	4.00	en cross	1.34	lasso bic	1.27	all subset best20% aic	0.16
5	all subset best20% aic	4.00	lasso cross	1.22	lasso cross	1.18	all subset best10% aic	0.16
6	all subset best30% aic	3.55	lars cross	1.22	lars cross	1.18	all subset best30% aic	0.15
7	all subset best aic	3.03	lars aic	0.93	en R^2	0.90	all subset best aic	0.12
8	en aic	2.30	lasso aic	0.93	en mse	0.89	en aic	0.09
9	30% across algorithms R^2	2.26	all subset best R^2	0.92	ℓ_2	0.84	30% across algorithms R^2	0.09
10	30% across algorithms mse	2.25	all subset best mse	0.91	lars aic	0.81	lars 10% R^2	0.09

Table 5.4.: Rankings of the first 10 best Linear Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	all subset best20% R^2	1.76	lasso R^2	0.61	lasso R^2	0.60	all subset best20% R^2	0.08
2	all subset best20% mse	1.76	lasso mse	0.61	lasso mse	0.60	all subset best20% mse	0.08
3	all subset best30% R^2	1.69	lasso aic	0.61	lasso aic	0.60	all subset best30% R^2	0.07
4	all subset best30% mse	1.69	lasso bic	0.61	lasso bic	0.60	all subset best30% mse	0.07
5	lasso cross	1.59	all subset best R^2	0.60	all subset best R^2	0.57	lasso cross	0.06
6	en 30% R^2	1.57	all subset best aic	0.60	all subset best aic	0.57	en 30% R^2	0.05
7	en 30% mse	1.57	all subset best bic	0.60	all subset best bic	0.57	en 30% mse	0.05
8	en 10% cross	1.56	all subset best mse	0.59	all subset best30% aic	0.57	en 10% cross	0.05
9	en 20% cross	1.56	all subset best30% aic	0.59	all subset best30% bic	0.57	en 20% cross	0.05
10	all subset best30% aic	1.52	all subset best30% bic	0.59	all subset best mse	0.56	10% across algorithms R^2	0.05

Table 5.5.: Rankings of asset allocation strategies for EQUITIES of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	lasso 30% aic	1.77	en 10% aic	0.20	en 10% aic	0.23	lasso 30% aic	0.20
2	lasso 30% bic	1.77	en 10% bic	0.20	en 10% bic	0.23	lasso 30% bic	0.20
3	all subset best10% R^2	1.69	en 20% aic	0.20	en 20% aic	0.23	all subset best10% R^2	0.18
4	all subset best10% mse	1.69	en 20% bic	0.20	en 20% bic	0.23	all subset best10% mse	0.18
5	en 10% cross	1.66	en 30% aic	0.20	en 30% aic	0.23	lasso BA	0.16
6	en 20% cross	1.66	en 30% bic	0.20	en 30% bic	0.23	all subset best30% R^2	0.15
7	all subset best30% R^2	1.65	spca	0.15	spca	0.16	all subset best30% mse	0.15
8	all subset best30% mse	1.65	lasso R^2	0.10	lasso 10% aic	0.11	en 10% cross	0.14
9	lasso BA	1.65	lasso mse	0.10	lasso 10% bic	0.11	en 20% cross	0.14
10	en 10% aic	1.64	lasso aic	0.10	lasso 20% aic	0.11	lasso 10% cross	0.14

Table 5.6.: Rankings of asset allocation strategies for BONDS of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	lasso 10% cross	3.31	lasso 30% aic	0.48	all subset best10% aic	0.33	30% across algorithms R^2	0.17
2	lasso 20% cross	3.31	lasso 30% bic	0.48	all subset best10% bic	0.33	30% across algorithms mse	0.17
3	all subset best20% aic	2.95	en cross	0.47	lasso 30% aic	0.30	all subset best20% aic	0.15
4	all subset best20% bic	2.95	ℓ_2 cross	0.47	lasso 30% bic	0.30	all subset best20% bic	0.15
5	lasso 30% cross	2.78	lasso 10% aic	0.47	en cross	0.29	10% across algorithms R^2	0.15
6	lasso 30% R^2	2.70	lasso 10% bic	0.47	ℓ_2 cross	0.29	10% across algorithms mse	0.15
7	lasso 30% mse	2.70	en 10% aic	0.47	lasso 10% aic	0.29	10% across algorithms cross	0.15
8	all subset best30% R^2	2.68	en 10% bic	0.47	lasso 10% bic	0.29	20% across algorithms R^2	0.15
9	30% across algorithms aic	2.65	ℓ_2 10% aic	0.47	en 10% aic	0.29	20% across algorithms mse	0.15
10	30% across algorithms bic	2.65	ℓ_2 10% bic	0.47	en 10% bic	0.29	20% across algorithms cross	0.15

Table 5.7.: Rankings of asset allocation strategies for COMMODITIES of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	lasso BA	1.46	en 10% aic	0.00	10% across algorithms cross	0.00	lasso BA	0.07
2	all subset best30% R^2	1.45	en 10% bic	0.00	20% across algorithms cross	0.00	all subset best30% R^2	0.06
3	all subset best30% mse	1.45	en 20% aic	0.00	30% across algorithms cross	0.00	all subset best30% mse	0.06
4	lasso 30% R^2	1.44	en 20% bic	0.00	10% across algorithms R^2	0.00	lasso 30% R^2	0.05
5	lasso 30% mse	1.44	en 30% aic	0.00	10% across algorithms mse	0.00	lasso 30% mse	0.05
6	lasso 10% R^2	1.41	en 30% bic	0.00	10% across algorithms aic	0.00	lasso 10% R^2	0.03
7	lasso 10% mse	1.41	10% across algorithms R^2	0.00	10% across algorithms bic	0.00	lasso 10% mse	0.03
8	lasso 20% R^2	1.41	10% across algorithms mse	0.00	20% across algorithms R^2	0.00	lasso 20% R^2	0.03
9	lasso 20% mse	1.41	10% across algorithms aic	0.00	20% across algorithms mse	0.00	lasso 20% mse	0.03
10	lasso 10% cross	1.41	10% across algorithms bic	0.00	20% across algorithms aic	0.00	lasso 10% cross	0.03

Table 5.8.: Rankings of asset allocation strategies for FOREIGN EXCHANGE of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restriction.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	all subset best20% R^2	2.94	lasso R^2	0.28	lasso R^2	0.24	all subset best20% R^2	0.17
2	all subset best20% mse	2.94	lasso mse	0.28	lasso mse	0.24	all subset best20% mse	0.17
3	all subset best30% R^2	2.66	lasso aic	0.28	lasso aic	0.24	all subset best30% R^2	0.15
4	all subset best30% mse	2.66	lasso bic	0.28	lasso bic	0.24	all subset best30% mse	0.15
5	all subset best R^2	2.21	all subset best R^2	0.25	all subset best R^2	0.19	all subset best R^2	0.11
6	all subset best aic	2.21	all subset best aic	0.25	all subset best aic	0.19	all subset best aic	0.11
7	all subset best bic	2.21	all subset best bic	0.25	all subset best bic	0.19	all subset best bic	0.11
8	all subset best30% aic	2.19	all subset best mse	0.24	all subset best30% aic	0.18	all subset best30% aic	0.11
9	all subset best30% bic	2.19	all subset best30% aic	0.23	all subset best30% bic	0.18	all subset best30% bic	0.11
10	lasso cross	2.18	all subset best30% bic	0.23	all subset best mse	0.17	lasso cross	0.11

Table 5.9.: Rankings of asset allocation strategies for EQUITIES of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	en 10% aic	2.05	en 10% aic	-0.07	en 10% aic	0.02	en 10% aic	0.13
2	en 10% bic	2.05	en 10% bic	-0.07	en 10% bic	0.02	en 10% bic	0.13
3	en 20% aic	2.05	en 20% aic	-0.07	en 20% aic	0.02	en 20% aic	0.13
4	en 20% bic	2.05	en 20% bic	-0.07	en 20% bic	0.02	en 20% bic	0.13
5	en 30% aic	2.05	en 30% aic	-0.07	en 30% aic	0.02	en 30% aic	0.13
6	en 30% bic	2.05	en 30% bic	-0.07	en 30% bic	0.02	en 30% bic	0.13
7	lasso 30% aic	2.00	spca	-0.18	spca	-0.12	lasso 30% aic	0.12
8	lasso 30% bic	2.00	lasso 10% aic	-0.26	lasso 10% aic	-0.21	lasso 30% bic	0.12
9	lasso 10% aic	1.71	lasso 10% bic	-0.26	lasso 10% bic	-0.21	lasso 10% aic	0.08
10	lasso 10% bic	1.71	lasso 20% aic	-0.26	lasso 20% aic	-0.21	lasso 10% bic	0.08

Table 5.10.: Rankings of asset allocation strategies for BONDS of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	lasso 10% cross	6.37	lasso 30% aic	0.48	all subset best10% aic	0.36	lasso 10% cross	0.21
2	lasso 20% cross	6.37	lasso 30% bic	0.48	all subset best10% bic	0.36	lasso 20% cross	0.21
3	lasso 30% cross	4.52	en cross	0.47	lasso 30% aic	0.30	lasso 30% cross	0.17
4	lasso 30% R^2	3.99	ℓ_2 cross	0.47	lasso 30% bic	0.30	lasso 30% R^2	0.16
5	lasso 30% mse	3.99	lasso 10% aic	0.47	en cross	0.29	lasso 30% mse	0.16
6	all subset best20% aic	3.58	lasso 10% bic	0.47	ℓ_2 cross	0.29	all subset best20% aic	0.15
7	all subset best20% bic	3.58	en 10% aic	0.47	lasso 10% aic	0.29	all subset best20% bic	0.15
8	spca	3.57	en 10% bic	0.47	lasso 10% bic	0.29	spca	0.14
9	all subset best30% aic	2.90	ℓ_2 10% aic	0.47	en 10% aic	0.29	all subset best30% aic	0.12
10	all subset best30% bic	2.90	ℓ_2 10% bic	0.47	en 10% bic	0.29	all subset best30% bic	0.12

Table 5.11.: Rankings of asset allocation strategies for COMMODITIES of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction.

Ranks	Method	Cum Wealth	Method	FF alphs	Method	Carhart alpha	Method	Sharpe Ratio
1	lasso BA	1.68	en 10% aic	0.09	en 10% aic	0.08	lasso BA	0.16
2	lasso 30% R^2	1.60	en 10% bic	0.09	en 10% bic	0.08	lasso 30% R^2	0.12
3	lasso 30% mse	1.60	en 20% aic	0.09	en 20% aic	0.08	lasso 30% mse	0.12
4	all subset best30% R^2	1.58	en 20% bic	0.09	en 20% bic	0.08	all subset best30% R^2	0.11
5	all subset best30% mse	1.58	en 30% aic	0.09	en 30% aic	0.08	all subset best30% mse	0.11
6	lasso 10% R^2	1.56	en 30% bic	0.09	en 30% bic	0.08	lasso 10% R^2	0.10
7	lasso 10% mse	1.56	lasso 10% aic	0.09	lasso 10% aic	0.08	lasso 10% mse	0.10
8	lasso 20% R^2	1.56	lasso 10% bic	0.09	lasso 10% bic	0.08	lasso 20% R^2	0.10
9	lasso 20% mse	1.56	lasso 20% aic	0.09	lasso 20% aic	0.08	lasso 20% mse	0.10
10	lasso 10% cross	1.55	lasso 20% bic	0.09	lasso 20% bic	0.08	lasso 10% cross	0.10

Table 5.12.: Rankings of asset allocation strategies for FOREIGN EXCHANGE of the first 10 best Logit Models and averaging schemes according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restriction.

	Method	Change Cum Wealth & %success prediction	Method	Change FF alphas & %success prediction	Method	Change Carhart alpha & %success prediction	Method	Change Sharpe Ratio & %success prediction
1	all subset best20% mse	1	10% across algorithms R^2	0	10% across algorithms R^2	0	20% across algorithms cross	0
2	all subset best30% mse	1	20% across algorithms mse	0	20% across algorithms mse	0	30% across algorithms cross	1
3	all subset best R^2	2	spca	0	lasso 30% cross	0	all subset best20% mse	1
4	en 20% mse	2	10% across algorithms mse	2	10% across algorithms mse	2	all subset best30% mse	1
5	lasso 10% cross	2	20% across algorithms R^2	2	20% across algorithms R^2	2	ℓ_2 30% aic	1
6	lasso 20% cross	2	30% across algorithms mse	2	30% across algorithms mse	2	ℓ_2 30% bic	1
7	all subset best20% R^2	3	all subset best aic	3	spca	2	10% across algorithms cross	2
8	all subset best30% R^2	3	all subset best bic	3	all subset best aic	3	en 20% mse	2

9	en 20% R^2	4	30% across algorithms R^2	4	all subset best bic	3	lasso 10% cross	2
10	lasso cross	4	30% across algorithms aic	5	30% across algorithms R^2	4	lasso 20% cross	2

Table 5.13.: Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restrictions for EQUITIES.

	Method	**Change Cum Wealth & %success prediction**	**Method**	**Change FF alphas & %success prediction**	**Method**	**Change Carhart alpha & %success prediction**	**Method**	**Change Sharpe Ratio & %success prediction**
1	20% across algorithms aic	0	en 10% mse	0	en mse	0	20% across algorithms R^2	1
2	20% across algorithms bic	0	en aic	0	20% across algorithms R^2	1	ℓ_2 BA	1
3	20% across algorithms cross	0	20% across algorithms R^2	1	ℓ_2 10% R^2	1	lasso 10% cross	1

4	30% across algorithms aic	0	all subset best10% mse	1	20% across algorithms aic	2	lasso 20% cross	1
5	30% across algorithms bic	0	en bic	1	20% across algorithms bic	2	lasso 30% cross	1
6	30% across algorithms cross	0	20% across algorithms aic	2	20% across algorithms cross	2	20% across algorithms aic	2
7	20% across algorithms mse	1	20% across algorithms bic	2	30% across algorithms aic	2	20% across algorithms bic	2
8	20% across algorithms R^2	1	20% across algorithms cross	2	30% across algorithms bic	2	20% across algorithms cross	2
9	ℓ_2 BA	1	30% across algorithms aic	2	30% across algorithms cross	2	30% across algorithms aic	2
10	10% across algorithms aic	2	30% across algorithms bic	2	en 10% aic	2	30% across algorithms bic	2

Table 5.14.: : Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restrictions for BONDS.

	Method	Change Cum Wealth & %success prediction	Method	Change FF alphas & %success prediction	Method	Change Carhart alpha & %success prediction	Method	Change Sharpe Ratio & %success prediction
1	en mse	0	20% across algorithms R^2	0	20% across algorithms R^2	0	en R^2	1
2	ℓ_2 10% R^2	1	30% across algorithms cross	0	30% across algorithms cross	0	lasso 10% R^2	1
3	lasso 10% R^2	1	ℓ_2 10% R^2	0	30% across algorithms mse	1	lasso aic	1
4	lasso aic	1	30% across algorithms mse	1	30% across algorithms R^2	1	lasso bic	1
5	lasso bic	1	30% across algorithms R^2	1	10% across algorithms R^2	2	lasso mse	2
6	en 10% cross	2	en 20% aic	1	20% across algorithms mse	2	en 30% mse	3
7	en 20% R^2	2	en 20% bic	1	en 10% aic	2	en mse	3
8	en R^2	2	10% across algorithms R^2	2	en 10% bic	2	lasso 10% mse	3
9	lasso mse	2	20% across algorithms mse	2	spca	2	ℓ_2 10% R^2	4

10	en aic	3	en 10% cross	2	all subset best20% R^2	3	lasso BA	4

Table 5.15.: : Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restrictions for COMMODITIES.

	Method	**Change Cum Wealth & %success prediction**	**Method**	**Change FF alphas & %success prediction**	**Method**	**Change Carhart alpha & %success prediction**	**Method**	**Change Sharpe Ratio & %success prediction**
1	en R^2	0	ℓ_2 30% cross	1	ℓ_2 30% cross	3	en R^2	0
2	lasso BA	0	ℓ_2 10% cross	2	lasso aic	7	lasso BA	0
3	ℓ_2 mse	1	ℓ_2 20% cross	2	lasso bic	7	lasso 30% mse	1
4	ℓ_2 R^2	1	ℓ_2 cross	5	ℓ_2 10% cross	10	en mse	2
5	lasso 30% mse	1	lasso aic	8	ℓ_2 20% cross	10	all subset best20% mse	3
6	en mse	2	lasso bic	8	lasso mse	10	lasso 30% R^2	3
7	all subset best20% mse	3	lasso mse	11	ℓ_2 BA	12	all subset best10% R^2	5

8	lasso 30% R^2	3	ℓ_2 R^2	12	lasso R^2	12	all subset best20% R^2	5
9	ℓ_2 aic	4	all subset best mse	13	ℓ_2 cross	13	en aic	5
10	ℓ_2 bic	4	lasso R^2	13	all subset best10% aic	14	en bic	5

Table 5.16.: : Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under Short Sales restrictions for FOREIGN EXCHANGE.

	Method	**Change Cum Wealth & %success prediction**	**Method**	**Change FF alphas & %success prediction**	**Method**	**Change Carhart alpha & %success prediction**	**Method**	**Change Sharpe Ratio & %success prediction**
1	30% across algorithms cross	0	10% across algorithms R^2	0	10% across algorithms R^2	0	30% across algorithms cross	0
2	en 10% cross	0	20% across algorithms mse	0	20% across algorithms mse	0	en 10% cross	0
3	ℓ_2 20% cross	0	spca	0	lasso 30% cross	0	ℓ_2 20% cross	0

4	10% across algorithms cross	1	10% across algorithms mse	2	spca	1	10% across algorithms cross	1
5	20% across algorithms cross	1	20% across algorithms R^2	2	10% across algorithms mse	2	20% across algorithms cross	1
6	all subset best20% aic	1	30% across algorithms mse	2	20% across algorithms R^2	2	all subset best20% aic	1
7	all subset best20% bic	1	all subset best aic	3	30% across algorithms mse	2	all subset best20% bic	1
8	all subset best20% mse	1	all subset best bic	3	all subset best aic	3	all subset best20% mse	1
9	all subset best30% mse	1	30% across algorithms R^2	4	all subset best bic	3	all subset best30% mse	1
10	lasso mse	1	30% across algorithms aic	5	30% across algorithms R^2	4	lasso mse	1

Table 5.17.: Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restrictions for EQUITIES.

	Method	Change Cum Wealth & %success prediction	Method	Change FF alphas & %success prediction	Method	Change Carhart alpha & %success prediction	Method	Change Sharpe Ratio & %success prediction
1	20% across algorithms aic	0	en aic	0	20% across algorithms R^2	1	20% across algorithms aic	0
2	20% across algorithms bic	0	20% across algorithms R^2	1	all subset best10% mse	1	20% across algorithms bic	0
3	20% across algorithms cross	0	en bic	1	en bic	1	20% across algorithms cross	0
4	30% across algorithms aic	0	20% across algorithms aic	2	ℓ_2 10% mse	1	30% across algorithms aic	0
5	30% across algorithms bic	0	20% across algorithms bic	2	ℓ_2 10% R^2	1	30% across algorithms bic	0
6	30% across algorithms cross	0	20% across algorithms cross	2	20% across algorithms aic	2	30% across algorithms cross	0
7	ℓ_2 20% cross	0	30% across algorithms aic	2	20% across algorithms bic	2	ℓ_2 20% cross	0
8	lasso 30% mse	0	30% across algorithms bic	2	20% across algorithms cross	2	lasso 30% mse	0

9	20% across algorithms R^2	1	30% across algorithms cross	2	30% across algorithms aic	2	20% across algorithms R^2	1
10	30% across algorithms mse	1	all subset best10% mse	2	30% across algorithms bic	2	30% across algorithms mse	1

Table 5.18.: Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restrictions for BONDS.

	Method	**Change Cum Wealth & %success prediction**	**Method**	**Change FF alphas & %success prediction**	**Method**	**Change Carhart alpha & %success prediction**	**Method**	**Change Sharpe Ratio & %success prediction**
1	10% across algorithms aic	1	20% across algorithms R^2	0	10% across algorithms R^2	0	10% across algorithms aic	1
2	10% across algorithms bic	1	30% across algorithms cross	0	20% across algorithms mse	0	10% across algorithms bic	1
3	ℓ_2 20% mse	1	30% across algorithms mse	1	30% across algorithms aic	1	ℓ_2 20% mse	1

4	ℓ_2 20% R^2	1	30% across algorithms R^2	1	30% across algorithms bic	1	ℓ_2 20% R^2	1
5	ℓ_2 30% cross	1	en 20% aic	1	30% across algorithms mse	1	ℓ_2 30% cross	1
6	lasso 30% aic	1	en 20% bic	1	30% across algorithms R^2	1	lasso 30% aic	1
7	lasso 30% bic	1	10% across algorithms R^2	2	10% across algorithms mse	2	lasso 30% bic	1
8	20% across algorithms R^2	2	20% across algorithms mse	2	20% across algorithms cross	2	20% across algorithms R^2	2
9	30% across algorithms cross	2	en 10% cross	2	20% across algorithms R^2	2	30% across algorithms cross	2
10	10% across algorithms R^2	4	ℓ_2 10% cross	2	30% across algorithms cross	2	10% across algorithms R^2	4

Table 5.19.: Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restrictions for COMMODITIES.

	Method	Change Cum Wealth & %success prediction	Method	Change FF alphas & %success prediction	Method	Change Carhart alpha & %success prediction	Method	Change Sharpe Ratio & %success prediction
1	20% across algorithms aic	0	ℓ_2 30% cross	1	ℓ_2 30% cross	4	20% across algorithms aic	0
2	20% across algorithms bic	0	lasso aic	1	spca	4	20% across algorithms bic	0
3	en mse	0	lasso bic	1	lasso aic	7	en mse	0
4	lasso BA	0	spca	3	lasso bic	7	lasso BA	0
5	10% across algorithms aic	1	ℓ_2 10% cross	4	lasso mse	10	10% across algorithms aic	1
6	10% across algorithms bic	1	ℓ_2 20% cross	4	ℓ_2 10% cross	11	10% across algorithms bic	1
7	30% across algorithms aic	1	lasso mse	4	ℓ_2 20% cross	11	30% across algorithms aic	1
8	30% across algorithms bic	1	lasso R^2	6	ℓ_2 BA	11	30% across algorithms bic	1
9	30% across algorithms cross	1	ℓ_2 cross	7	lasso R^2	12	30% across algorithms cross	1

10	en aic	1	all subset best mse	13	ℓ_2 cross	14	en aic	1	

Table 5.20.: Table of the change in the rankings when comparing the percentage of successful prediction of the direction using Logit regression based Models and averaging schemes and the success when performing asset allocation strategies according to Cum wealth, Fama French alpha, Curhart alpha and Sharpe Ratio under no Short Sales restrictions for FOREIGN EXCHANGE.

EQUITY

Rank	**Model**	**Times one model predicts better than others (95% conf. level)**	**Rank**	**Model**	**Times one model predicts better than others (95% conf. level)**
1	all subset_best R^2	79	31	lasso 20% mse	4
2	all subset_best mse	79	32	lasso 20% aic	4
3	all subset_best aic	57	33	lasso 20% bic	3
4	all subset_best bic	55	34	lars 30% R^2	3
5	all subset_best10% R^2	54	35	lars 30% mse	3
6	all subset_best10% mse	54	36	lars 30% aic	3
7	all subset_best20% R^2	53	37	lars 30% bic	3
8	all subset_best20% mse	51	38	lasso 30% R^2	3
9	all subset_best20% aic	22	39	lasso 30% mse	3
10	all subset_best20% bic	18	40	lasso 30% aic	3

11	all subset_best30% R^2	16	41	lasso 30% bic	2
12	all subset_best30% mse	15	42	lars 10% cross	2
13	all subset_best30% aic	14	43	lasso 10% cross	2
14	all subset_best30% bic	12	44	en 10% cross	2
15	lars R^2	12	45	lars 20% cross	2
16	lars mse	8	46	lasso 20% cross	2
17	lars aic	8	47	en 20% cross	2
18	lars bic	7	48	lars 30% cross	2
19	lars cross	6	49	lasso 30% cross	2
20	lasso R^2	5	50	en 30% cross	2
21	lasso mse	5	51	10% across algorithms R^2	2
22	lasso bic	5	52	10% across algorithms mse	1
23	lars 10% R^2	5	53	10% across algorithms aic	1
24	lars 10% mse	5	54	10% across algorithms cross	1
25	lars 10% aic	5	55	20% across algorithms mse	1
26	lasso 10% R^2	5	56	20% across algorithms aic	1
27	lasso 10% mse	4	57	20% across algorithms cross	1
28	lasso 10% aic	4	58	30% across algorithms R^2	1
29	lars 20% aic	4	59	30% across algorithms mse	1
30	lars 20% bic	4			

Table 5.21.: Diebold-Mariano test for EQUITIES at 95% confidence level.

BONDS

Rank	Model	Times one model predicts better than others (95% conf. level)	Rank	Model	Times one model predicts better than others (95% conf. level)
1	all subset_best R^2	60	31	lasso 10% mse	12
2	all subset_best mse	59	32	lars 20% R^2	12
3	all subset_best aic	57	33	lars 20% mse	12
4	all subset_best bic	56	34	lasso 20% R^2	11
5	all subset_best10% R^2	54	35	lasso 20% mse	9
6	all subset_best10% mse	52	36	lars 30% R^2	9
7	all subset_best10% aic	46	37	lars 30% mse	9
8	all subset_best10% bic	46	38	lars 30% bic	9
9	all subset_best20% R^2	34	39	lasso 30% R^2	9
10	all subset_best20% mse	21	40	lasso 30% mse	9
11	all subset_best20% aic	20	41	lasso 30% bic	9
12	all subset_best20% bic	20	42	lars 10% cross	9
13	all subset_best30% R^2	19	43	lasso 10% cross	9
14	all subset_best30% mse	18	44	en 10% cross	9
15	all subset_best30% aic	18	45	lars 20% cross	7
16	all subset_best30% bic	18	46	lasso 20% cross	6
17	lars R^2	18	47	en 20% cross	5
18	lars mse	18	48	lars 30% cross	5

19	lars aic	18	49	lasso 30% cross	5
20	lars bic	18	50	en 30% cross	4
21	lars cross	17	51	10% across algorithms R^2	4
22	lasso R^2	16	52	10% across algorithms mse	3
23	lasso mse	16	53	10% across algorithms cross	3
24	lasso cross	15	54	20% across algorithms R^2	3
25	en R^2	15	55	20% across algorithms mse	3
26	en mse	15	56	20% across algorithms cross	2
27	en aic	15	57	30% across algorithms R^2	2
28	lars 10% R^2	15	58	30% across algorithms mse	2
29	lars 10% mse	14	59	30% across algorithms bic	1
30	lasso 10% R^2	14	60	30% across algorithms cross	1

Table 5.22.: Diebold-Mariano test for BONDS at 95% confidence level.

COMMODITY

Rank	**Model**	**Times one model predicts better than others (95% conf. level)**	**Rank**	**Model**	**Times one model predicts better than others (95% conf. level)**
1	all subset_best R^2	82	33	en 10% bic	9
2	all subset_best mse	82	34	lars 20% R^2	9
3	all subset_best10% R^2	60	35	lars 20% mse	8

4	all subset_best10% mse	60	36	lars 20% bic	8
5	all subset_best10% aic	58	37	lasso 20% R^2	8
6	all subset_best10% bic	58	38	lasso 20% mse	8
7	all subset_best20% R^2	28	39	lasso 20% bic	6
8	all subset_best20% mse	28	40	en 20% R^2	6
9	all subset_best20% aic	25	41	en 20% mse	6
10	all subset_best20% bic	22	42	en 20% bic	4
11	all subset_best30% R^2	22	43	lars 30% R^2	4
12	all subset_best30% mse	21	44	lars 30% mse	3
13	all subset_best30% aic	21	45	lars 30% bic	3
14	all subset_best30% bic	21	46	lasso 30% R^2	3
15	lars R^2	19	47	lasso 30% mse	3
16	lars mse	18	48	lasso 30% bic	3
17	lars aic	18	49	en 30% R^2	3
18	lasso R^2	17	50	en 30% mse	3
19	lasso mse	16	51	lars 10% cross	2
20	lasso aic	15	52	lasso 10% cross	1
21	spca	15	53	en 10% cross	1
22	lars 10% R^2	14	54	lars 30% cross	1
23	lars 10% mse	13	55	lasso 30% cross	1
24	lars 10% aic	13	56	en 30% cross	1
25	lars 10% bic	12	57	10% across algorithms R^2	1

26	lasso 10% R^2	12	58	10% across algorithms mse	1
27	lasso 10% mse	12	59	10% across algorithms aic	1
28	lasso 10% aic	11	60	20% across algorithms R^2	1
29	lasso 10% bic	10	61	20% across algorithms mse	1
30	en 10% R^2	10	62	30% across algorithms R^2	1
31	en 10% mse	10	63	30% across algorithms mse	1
32	en 10% aic	10			

Table 5.23.: Diebold-Mariano test for COMMODITIES at 95% confidence level.

Foreign Exchange

Rank	**Model**	**Times one model predicts better than others (95% conf. level)**	**Rank**	**Model**	**Times one model predicts better than others (95% conf. level)**
1	all subset_best R^2	74	32	en 10% aic	3
2	all subset_best mse	74	33	en 10% bic	3
3	all subset_best aic	61	34	lars 20% R^2	3
4	all subset_best bic	40	35	lars 20% mse	3
5	all subset_best10% R^2	35	36	lasso 20% R^2	3
6	all subset_best10% mse	34	37	lasso 20% mse	3
7	all subset_best10% bic	29	38	en 20% R^2	3
8	all subset_best20% R^2	28	39	en 20% mse	3

9	all subset_best20% mse	28	40	en 20% aic	3
10	all subset_best20% bic	25	41	en 20% bic	3
11	all subset_best30% R^2	25	42	lars 30% R^2	3
12	all subset_best30% mse	16	43	lars 30% mse	3
13	lars R^2	15	44	lars 30% bic	3
14	lars mse	11	45	lasso 30% R^2	3
15	lars aic	11	46	lasso 30% mse	3
16	lasso R^2	9	47	lasso 30% bic	2
17	lasso mse	9	48	en 30% R^2	2
18	lasso aic	8	49	en 30% mse	2
19	en R^2	7	50	en 30% aic	1
20	en mse	7	51	en 30% bic	1
21	en aic	7	52	lars 30% cross	1
22	en bic	5	53	lasso 30% cross	1
23	spca	5	54	en 30% cross	1
24	lars 10% R^2	4	55	10% across algorithms R^2	1
25	lars 10% mse	4	56	10% across algorithms mse	1
26	lars 10% bic	4	57	10% across algorithms bic	1
27	lasso 10% R^2	3	58	20% across algorithms R^2	1
28	lasso 10% mse	3	59	20% across algorithms mse	1
29	lasso 10% bic	3	60	30% across algorithms R^2	1
30	en 10% R^2	3	61	30% across algorithms mse	1

31	en 10% mse	3

Table 5.24.: Diebold-Mariano test for FOREIGN EXCHANGE at 95% confidence level.

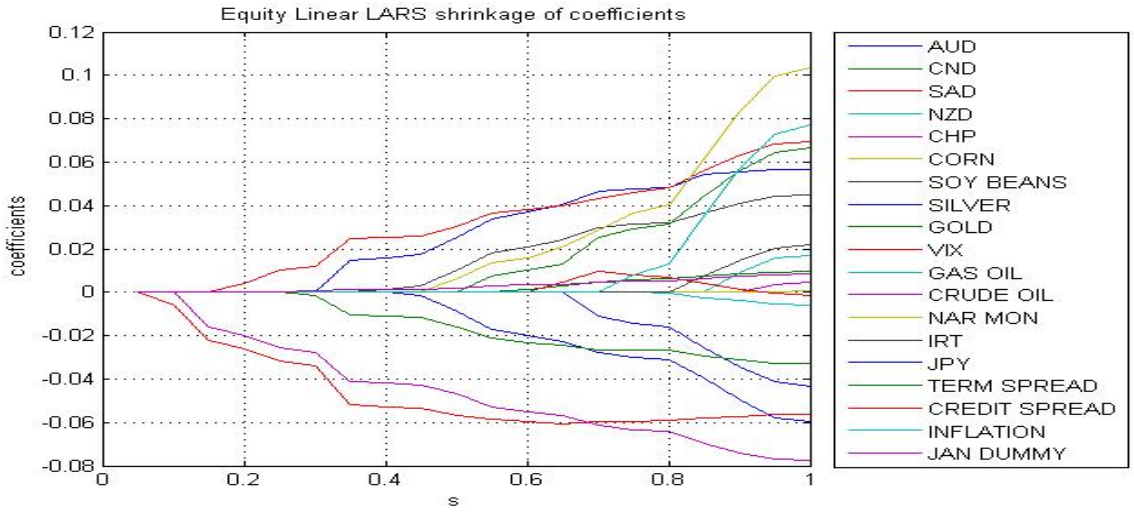


Figure 5.1.: LARS shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the linear regression application on equities. Each line represents a coefficient.

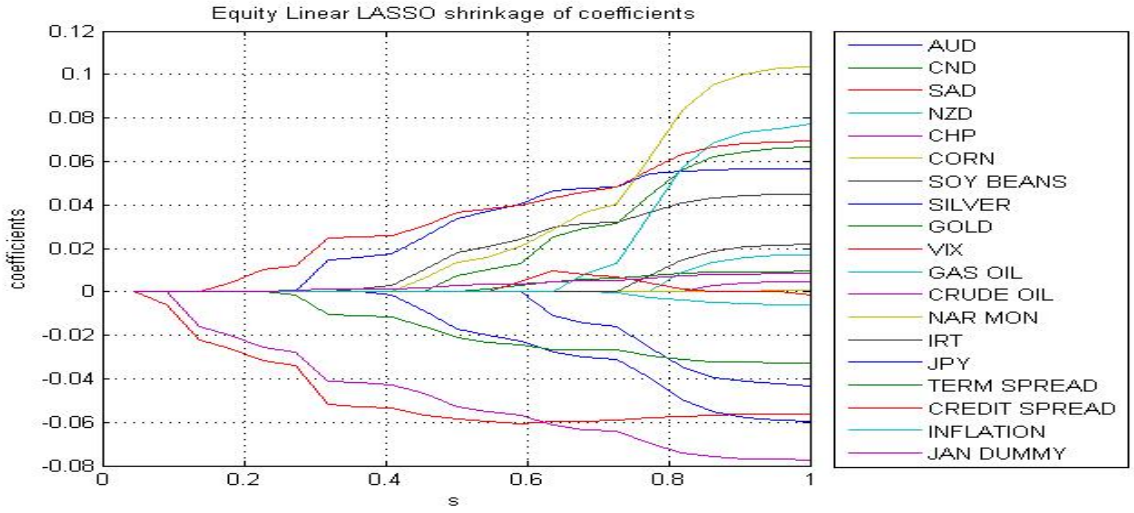


Figure 5.2.: LASSO shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the linear regression application on equities.

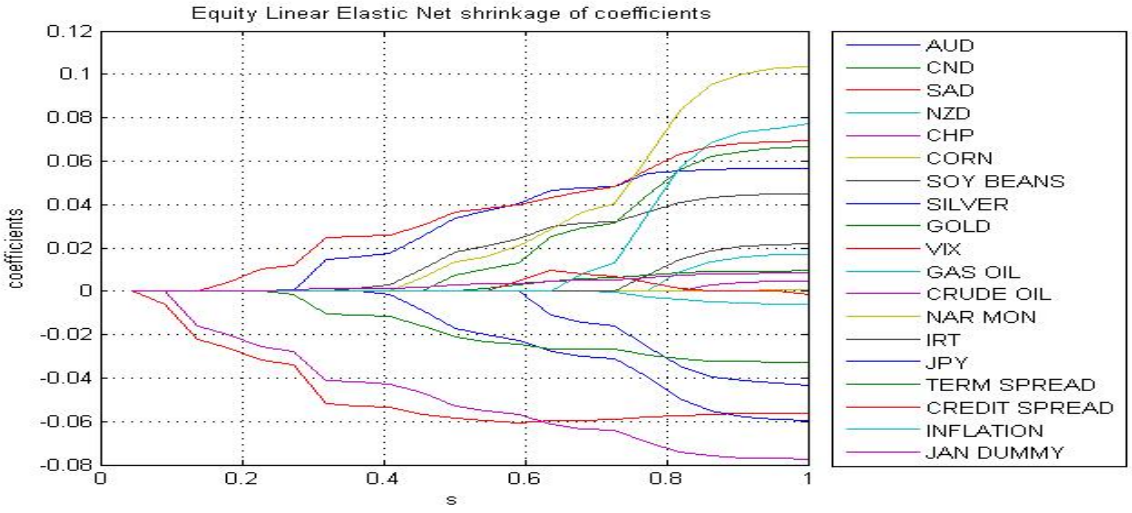


Figure 5.3.: Elastic Net shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the linear regression application on equities.

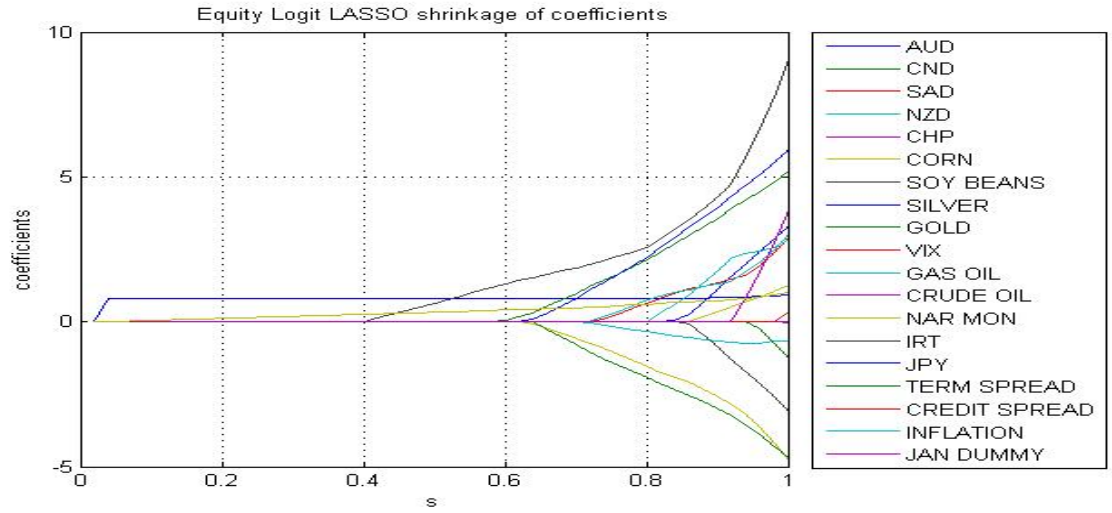


Figure 5.4.: LASSO shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the logit regression application on equities.

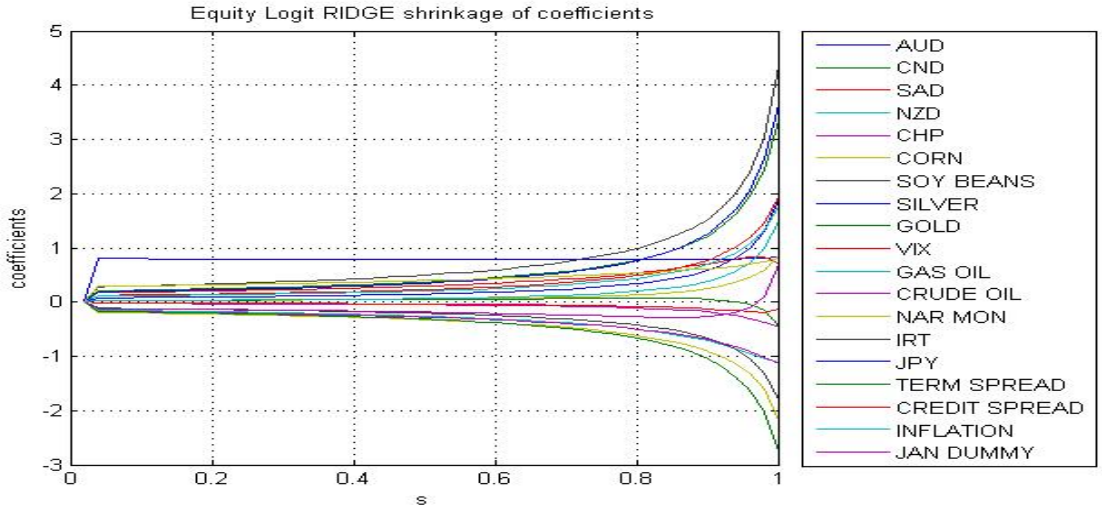


Figure 5.5.: Ridge regression shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the logit regression application on equities.

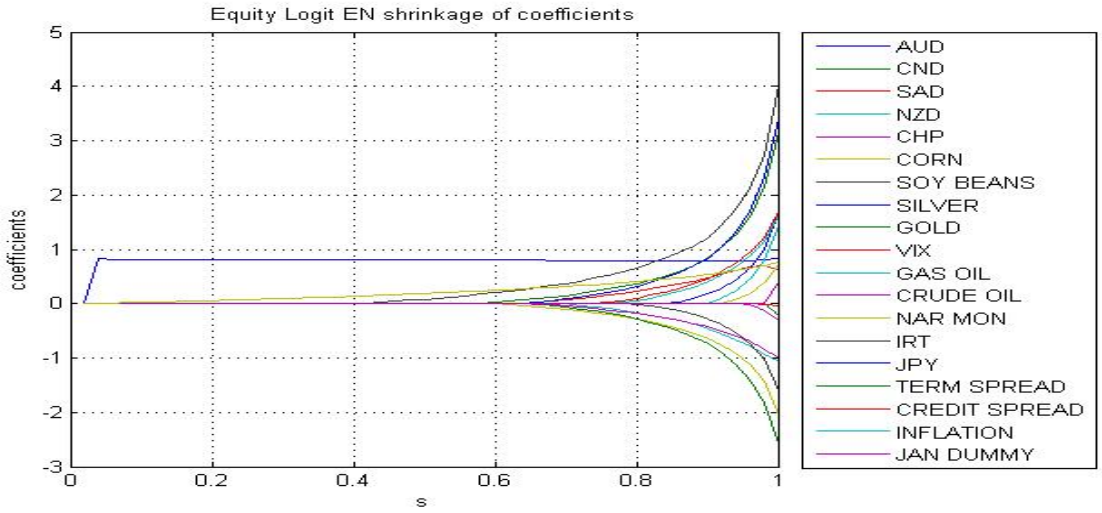


Figure 5.6.: Elastic Net shrinkage of coefficients as a function of parameter $s = \frac{\beta}{\sum|\hat{\beta}_j|}$ for the logit regression application on equities.

6. Conclusions & Further Research

This chapter summarizes the most important conclusions derived from the thesis, and sets up a framework for future improvements and extensions of all methods used in various problem settings.

From what has been discussed so far, we can conclude that there is no clear indication of the dominance of a single forecasting model that could always generate the same outperformance in different scenarios, market environments and time periods. The subjectivity inherent in the researcher's choice of the model selection criterion which is used to indicate superiority of a single process against other alternatives during the estimation period unavoidably creates biases. We have proved empirically that this selection based on various econometric criteria fails to produce the desired outcome for an extended targeted forecasting period; thus the uncertainty that already exists escalates as we move further into the future to make predictions. To address all these issues, model-averaging approaches are proposed. Their implementation has been extended here from forecasting moments to predicting distributions, pricing options and improving the performance of complicated model-selection algorithms. This is in addition to their application in portfolio-related investment decisions, where they facilitate in the capital allocation process in order to generate excess wealth.

These are just some among many examples where model averaging can be utilized, since the limitations of the method are few. One of those limitations, however, relates to the fact that most of these averaging routines can be very expensive computationally. The use of a supercomputer in many instances was unavoidable, so the researcher needs sufficient programming experience and the tools to execute these advanced computational problems. However, some of the methods, such as model selection algorithms encountered in the fifth chapter, can be executed quickly and their performance can be improved under the averaging spectrum. In other words, it is not always the case that averaging creates an additional burden: it may also lift

some barriers without significantly exploding the estimation time. Also the scale of the problems and the targeted time horizons that were employed for forecasting purposes in the thesis were to say the least extensive. In reality, predicting just one step ahead at a time is a sufficient forecasting time-frame, which therefore can be plausibly approached using averaging techniques.

In addition, regarding future steps we have to notice the following. All models discussed here were under the univariate spectrum. Extending the averaging suggestions to the multivariate case where correlation dynamics play a significant role might also be a challenge. By definition, all multivariate models have parameters of higher order, thus estimating a collection of them to perform averaging might explode the total number of coefficient-related calculations. However, imposing restrictions on their specifications could facilitate generating more sparse solutions. It is also suggested that an investigation related to alternative optimisation routines, since those employed have been the core of all research chapters, could vastly improve computational time and accurateness. For the time being, however, this was beyond the scope of the thesis. But in an era where computers approach the teraflops region, the prospects of averaging methods seem better than ever before.

Regarding model selection algorithms, the generalization of them was performed under the logit regression framework. Additional loss functions could also be employed in order to investigate the performance of the averaging methods and ultimately gain a wider perspective of their usefulness. This could also offer a better insight regarding which algorithms work better for which type of data and which transformations. Further work is also suggested when some variables are forced to be included into the model while the rest are not. In other words, it might happen that the investor has firm beliefs that some variables are important to describe the movement of the dependent variable. Thus these should definitely be included in the predictive equation. It is left to be investigated how model selection algorithms react to such restrictions and whether performance is indeed improved or not under certain constraints.

Additional work is also needed regarding the tuning parameter. Although this has been also discussed in the past it has never been investigated under the averaging spectrum. Usual criteria include cross-validation, AIC, BIC,

C_p for LARS etc, the usefulness of which may depend on the averaging method as well as on the data or the model selection algorithm picked.

Finally, regarding the type of data used in all chapters we should notice the following. Although the main focus was equity indexes, further exploration is suggested in other types of assets. This is in order to substantiate a wider acceptance of the superiority of an averaging method against other alternatives. However, one cannot disregard that such an initiation was established in the fifth chapter, where in addition to equity we investigated bond, FX and commodity index forecastability. The predictors used for this purpose included commodity prices and exchange rates as well as commonly used macroeconomic indicators. What we set out to prove was that there are some global factors that affect highly tradable indexes. However, the list of such predictors cannot be exhausted, although results endorse our initial statement, i.e. that exchange rates and commodities can affect the performance of other asset classes. Further investigation is suggested, however, in order to validate such a strong assumption.

A. Additional Tables Chapter 3

min	-0.0711
max	0.0557
mean	0.0002
median	0.0001
var	0.0001
std	0.0112
kurt	6.3894
skew	-0.0843
jb	1
ks	1

Table A.1.: Summary statistics of S&P 500 for the period 31_{st} August 1997 to 1_{st} September 2007.

The above table contains some indicative statistics for the S&P 500 index for the years starting from 31/8/1997 until 1/9/2007 (a total of 2610 observations). These statistics refer to the whole sample (both in and out-of-sample). The last two rows refer to Bera-Jarque hypothesis test of normality and Kolmogorov-Smirnov goodness-of-fit hypothesis test. At significance level 0.05 the first test rejects the null hypothesis of normally distributed data. The same holds for the second test where the null hypothesis that the sample's CDF is normal is again rejected. These results are not surprising, since the above mentioned tests use the third and forth moment to determine whether the sample has the desired properties, but clearly these moments differ from their theoretical normal values which are three and zero respectively.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0	3.32E-09	TM 20% OPT	64	3	3	3.78E-09	NAGARCH	127	4	4	4.10E-09	NGARCH
2	1	2	3.33E-09	TGARCH	65	4	3	3.78E-09	NAGARCH	128	2	4	4.11E-09	FATGARCH
3	2	2	3.33E-09	TGARCH	66	1	1	3.79E-09	NGARCH	129	2	4	4.11E-09	GARCH
4	3	2	3.34E-09	TGARCH	67	3	3	3.79E-09	NGARCH	130	3	4	4.14E-09	GJRGARCH
5	2	1	3.34E-09	TGARCH	68	1	3	3.79E-09	NGARCH	131	1	4	4.15E-09	GJRGARCH
6	3	1	3.34E-09	TGARCH	69	4	4	3.79E-09	NAGARCH	132	2	4	4.15E-09	GJRGARCH
7	1	1	3.34E-09	TGARCH	70	4	4	3.80E-09	NGARCH	133	4	4	4.16E-09	GJRGARCH
8	4	2	3.34E-09	TGARCH	71	3	2	3.81E-09	NGARCH	134	0	0	4.32E-09	TM 50% NN
9	4	4	3.35E-09	TGARCH	72	3	4	3.81E-09	NGARCH	135	0	0	4.57E-09	TM 40% NN
10	4	1	3.35E-09	TGARCH	73	2	4	3.81E-09	NGARCH	136	4	0	5.81E-09	FATGARCH
11	0	0	3.36E-09	TM 10% OPT	74	1	1	3.82E-09	EGARCH	137	4	0	5.81E-09	GARCH
12	2	3	3.36E-09	TGARCH	75	3	3	3.82E-09	NGARCH	138	0	0	6.19E-09	TM 20% NN
13	1	3	3.36E-09	TGARCH	76	3	4	3.83E-09	NAGARCH	139	0	0	6.27E-09	TM 30% NN
14	3	3	3.36E-09	TGARCH	77	1	3	3.85E-09	GARCH	140	3	0	6.69E-09	FATGARCH
15	0	0	3.36E-09	TM	78	1	3	3.85E-09	FATGARCH	141	3	0	6.69E-09	GARCH
16	1	4	3.37E-09	TGARCH	79	1	2	3.85E-09	GARCH	142	1	0	6.93E-09	APGARCH
17	4	3	3.37E-09	TGARCH	80	1	2	3.85E-09	FATGARCH	143	1	2	6.93E-09	APGARCH
18	2	4	3.37E-09	TGARCH	81	4	1	3.85E-09	GARCH	144	2	0	6.93E-09	APGARCH
19	3	4	3.40E-09	TGARCH	82	4	1	3.85E-09	FATGARCH	145	2	2	6.93E-09	APGARCH
20	0	0	3.44E-09	TM 30% OPT	83	1	4	3.86E-09	GARCH	146	3	0	6.93E-09	APGARCH
21	0	0	3.44E-09	TM 30% OPT	84	1	4	3.86E-09	FATGARCH	147	3	2	6.93E-09	APGARCH

22	0	0	3.46E-09	TM 50%	85	1	4	3.87E-09	FIGARCH	148	3	3	6.93E-09	APGARCH
23	0	0	3.46E-09	TM 50% OPT	86	3	1	3.87E-09	GARCH	149	4	0	6.93E-09	APGARCH
24	2	1	3.47E-09	APGARCH	87	3	1	3.87E-09	FATGARCH	150	4	2	6.93E-09	APGARCH
25	0	0	3.48E-09	TM 40% OPT	88	1	1	3.88E-09	GARCH	151	1	0	6.93E-09	NGARCH
26	0	0	3.48E-09	TM 40%	89	1	1	3.88E-09	FATGARCH	152	2	0	6.93E-09	NGARCH
27	0	0	3.48E-09	TM 10%	90	4	2	3.90E-09	GARCH	153	3	0	6.93E-09	NGARCH
28	1	1	3.51E-09	APGARCH	91	4	2	3.90E-09	FATGARCH	154	4	0	6.93E-09	NGARCH
29	3	1	3.51E-09	APGARCH	92	2	2	3.90E-09	NGARCH	155	1	0	6.93E-09	NGARCH
30	2	1	3.56E-09	NAGARCH	93	2	3	3.90E-09	NAGARCH	156	2	0	6.93E-09	NGARCH
31	1	2	3.56E-09	NAGARCH	94	2	3	3.90E-09	NGARCH	157	3	0	6.93E-09	NGARCH
32	4	1	3.59E-09	APGARCH	95	3	4	3.90E-09	NGARCH	158	4	0	6.93E-09	NGARCH
33	3	1	3.60E-09	NAGARCH	96	2	1	3.93E-09	FATGARCH	159	1	0	6.93E-09	NAGARCH
34	0	0	3.61E-09	BA	97	2	1	3.93E-09	GARCH	160	2	0	6.93E-09	NAGARCH
35	1	1	3.61E-09	NGARCH	98	3	2	3.93E-09	GARCH	161	3	0	6.93E-09	NAGARCH
36	1	2	3.62E-09	NGARCH	99	3	2	3.93E-09	FATGARCH	162	4	0	6.93E-09	NAGARCH
37	3	1	3.62E-09	NGARCH	100	4	3	3.94E-09	FATGARCH	163	1	0	6.93E-09	TGARCH
38	2	1	3.62E-09	NGARCH	101	4	3	3.94E-09	GARCH	164	2	0	6.93E-09	TGARCH
39	3	2	3.63E-09	NAGARCH	102	1	3	3.96E-09	FIGARCH	165	3	0	6.93E-09	TGARCH
40	1	1	3.63E-09	NAGARCH	103	2	2	3.97E-09	FATGARCH	166	4	0	6.93E-09	TGARCH
41	1	3	3.63E-09	NAGARCH	104	2	2	3.97E-09	GARCH	167	1	0	6.93E-09	GJRGARCH
42	2	2	3.63E-09	NAGARCH	105	4	2	3.97E-09	GJRGARCH	168	2	0	6.93E-09	GJRGARCH
43	4	2	3.63E-09	NAGARCH	106	2	2	3.97E-09	GJRGARCH	169	3	0	6.93E-09	GJRGARCH

44	1	4	3.64E-09	NAGARCH	107	1	2	3.97E-09	GJRGARCH	170	4	0	6.93E-09	GJRGARCH
45	1	4	3.65E-09	NGARCH	108	3	1	3.97E-09	GJRGARCH	171	1	4	7.30E-09	APGARCH
46	4	2	3.65E-09	NGARCH	109	4	4	3.97E-09	FATGARCH	172	0	0	7.63E-09	TM 10% NN
47	4	1	3.65E-09	NAGARCH	110	4	4	3.98E-09	GARCH	173	4	4	7.78E-09	APGARCH
48	2	2	3.65E-09	NGARCH	111	1	1	3.98E-09	GJRGARCH	174	1	3	7.89E-09	APGARCH
49	3	2	3.68E-09	NGARCH	112	2	1	3.98E-09	GJRGARCH	175	2	3	8.09E-09	APGARCH
50	1	1	3.69E-09	FIGARCH	113	4	1	3.98E-09	GJRGARCH	176	2	0	8.25E-09	FATGARCH
51	0	0	3.69E-09	BMA	114	3	3	3.98E-09	FATGARCH	177	2	0	8.25E-09	GARCH
52	3	1	3.69E-09	NGARCH	115	3	3	3.98E-09	GARCH	178	3	4	8.29E-09	APGARCH
53	1	2	3.70E-09	FIGARCH	116	4	3	3.99E-09	NGARCH	179	4	3	8.51E-09	APGARCH
54	4	1	3.71E-09	NGARCH	117	3	2	4.00E-09	GJRGARCH	180	1	0	9.70E-09	FATGARCH
55	1	2	3.71E-09	NGARCH	118	2	4	4.00E-09	NGARCH	181	1	0	9.70E-09	GARCH
56	4	2	3.72E-09	NGARCH	119	2	3	4.02E-09	FATGARCH	182	2	4	1.06E-08	APGARCH
57	1	3	3.73E-09	NGARCH	120	2	3	4.02E-09	GARCH	183	0	1	1.20E-08	FATGARCH
58	2	4	3.74E-09	NAGARCH	121	3	4	4.03E-09	GARCH	184	0	2	1.20E-08	FATGARCH
59	4	1	3.77E-09	NGARCH	122	3	4	4.03E-09	FATGARCH	185	0	3	1.20E-08	FATGARCH
60	2	3	3.77E-09	NGARCH	123	4	3	4.06E-09	GJRGARCH	186	0	4	1.20E-08	FATGARCH
61	4	3	3.78E-09	NGARCH	124	3	3	4.07E-09	GJRGARCH	187	1	0	1.21E-08	FIGARCH
62	2	1	3.78E-09	NGARCH	125	1	3	4.07E-09	GJRGARCH					
63	1	4	3.78E-09	NGARCH	126	2	3	4.09E-09	GJRGARCH					

Table A.2.: Rankings of all volatility models and averaging methods according to RMSE.

Rank	p	q	MSE	Model	Rank	p	q	MSE	Model	Rank	p	q	MSE	Model
1	0	0	1.8347	TM 10% OPT	43	3	3	2.0579	NAGARCH	85	4	4	2.9652	NGARCH
2	0	0	1.845	TM 20% OPT	44	2	2	2.0587	NAGARCH	86	2	4	2.9963	NGARCH
3	0	0	1.8533	TM 30% OPT	45	1	4	2.0709	NAGARCH	87	4	3	3.0315	FATGARCH
4	0	0	1.8619	TM 40% OPT	46	3	3	2.077	NGARCH	88	2	4	3.0866	FATGARCH
5	0	0	1.8708	TM 50% OPT	47	0	0	2.0773	TM 10% VOL	89	4	4	3.1105	FATGARCH
6	3	1	1.8993	NGARCH	48	3	2	2.0814	NAGARCH	90	3	4	3.1148	NGARCH
7	2	1	1.8994	NAGARCH	49	2	1	2.0826	GARCH	91	2	3	3.1817	FATGARCH
8	2	2	1.9024	NGARCH	50	2	4	2.0869	NGARCH	92	3	3	3.2925	FATGARCH
9	2	1	1.9066	NGARCH	51	4	3	2.0907	NGARCH	93	4	2	3.3719	FATGARCH
10	0	0	1.9097	TM 30% OPT VOL	52	1	2	2.0979	GARCH	94	4	2	3.4622	GJRGARCH
11	1	1	1.9108	NGARCH	53	1	4	2.1002	NGARCH	95	4	1	3.5579	FATGARCH
12	0	0	1.9136	TM 40% OPT	54	4	1	2.1068	GARCH	96	3	4	3.6345	FATGARCH
13	0	0	1.9149	BA VOL	55	4	4	2.1074	NGARCH	97	3	2	3.6635	FATGARCH
14	0	0	1.915	BA OPT VOL	56	1	2	2.1188	NGARCH	98	2	2	3.7979	FATGARCH
15	0	0	1.919	TM 50%	57	1	3	2.1426	GARCH	99	3	1	3.8785	FATGARCH
16	0	0	1.9193	TM 20% OPT Vol	58	1	1	2.1456	NGARCH	100	1	1	3.9439	FATGARCH
17	0	0	1.9199	TM 40% OPT VOL	59	3	2	2.1622	NGARCH	101	1	2	4.0627	FATGARCH
18	0	0	1.9217	TM 50% OPT VOL	60	3	1	2.2092	GARCH	102	1	4	4.143	FATGARCH
19	1	2	1.9257	NAGARCH	61	2	2	2.215	NGARCH	103	1	3	4.204	FATGARCH
20	1	2	1.9272	NGARCH	62	1	3	2.2377	NGARCH	104	2	1	4.2148	FATGARCH
21	1	1	1.9276	NAGARCH	63	1	4	2.2394	GARCH	105	0	0	4.2963	BMA
22	0	0	1.9285	TM 10% OPT Vol	64	3	4	2.2639	NGARCH	106	3	1	5.3077	GJRGARCH
23	1	3	1.9301	NGARCH	65	3	1	2.3107	NGARCH	107	0	0	5.3178	BA

24	0	0	1.931	TM 30%	66	3	2	2.3353	GARCH	108	2	4	5.4172	GJRGARCH
25	0	0	1.9334	TM 20%	67	3	4	2.3361	NAGARCH	109	2	1	5.6509	GJRGARCH
26	0	0	1.9448	BMA VOL	68	3	2	2.3371	NGARCH	110	4	4	6.1269	GJRGARCH
27	3	1	1.9521	NAGARCH	69	4	2	2.3411	NGARCH	111	4	3	6.4497	GJRGARCH
28	1	3	1.9612	NAGARCH	70	2	4	2.3464	NAGARCH	112	1	4	6.8503	GJRGARCH
29	2	2	1.9816	GARCH	71	4	4	2.3495	GARCH	141	4	3	38.1304	TGARCH
30	2	3	1.9817	NAGARCH	72	3	3	2.3556	GARCH	142	3	4	42.7133	GJRGARCH
31	0	0	1.9864	TM 50% VOL	73	1	4	2.3748	NGARCH	143	4	4	47.1188	TGARCH
32	0	0	1.9939	TM 10%	74	4	1	2.375	NGARCH	144	1	2	48.5317	TGARCH
33	4	1	2.0047	NAGARCH	75	4	4	2.3944	NAGARCH	145	2	3	50.3563	GJRGARCH
34	0	0	2.0084	TM 40% VOL	76	2	4	2.4657	GARCH	146	2	2	55.5346	GJRGARCH
35	4	1	2.0128	NGARCH	77	4	2	2.5227	NGARCH	147	1	3	57.9269	GJRGARCH
36	2	3	2.0243	NGARCH	78	3	4	2.5609	GARCH	148	4	2	58.1431	TGARCH
37	1	1	2.0246	GARCH	79	2	3	2.6911	GARCH	149	1	3	75.0772	TGARCH
38	4	2	2.0319	NAGARCH	80	4	3	2.723	GARCH	150	3	1	76.6269	TGARCH
39	2	1	2.0358	NGARCH	81	2	3	2.7589	NGARCH	151	1	2	88.942	GJRGARCH
40	4	3	2.0368	NAGARCH	82	3	3	2.7621	NGARCH	152	2	3	104.4589	TGARCH
41	0	0	2.0429	TM 30% VOL	83	4	2	2.8523	GARCH	153	2	2	180.1334	TGARCH
42	0	0	2.0538	TM 20% VOL	84	4	3	2.9568	NGARCH	154	4	1	185.5587	TGARCH

Table A.3.: Rankings according to $RMSE of all option pricing models and Model Averaging methods.

B. Additional Tables Chapter 5

	Category	Name	Description
1	FX	US Dollar Index	Index of the value of the United States dollar relative to a basket of foreign currencies. It is a weighted geometric mean of the dollar's value compared to the euro (EUR), Japanese yen (JPY), Pound sterling (GBP), Canadian dollar (CAD), Swedish krona (SEK) and Swiss franc (CHF)
2	FX	USD/AUD	exchange rate Australian dollar against the U.S. dollar
3	FX	USD/CAD	exchange rate Canadian dollar against the U.S. dollar
4	FX	USD/ZAR	exchange rate South African Rand against the U.S. dollar
5	FX	USD/NZD	exchange rate New Zealand dollar against the U.S. dollar
6	FX	USD/CLP	exchange rate Chilean Peso against the U.S. dollar
7	FX	USD/JPY	

8	COMM	CORN: #2 YELLOW(MAIZE): 1ST EXPIRATION FUTURE NEARBY - SETTLEMENT - CHICAGO BOARD OF TRADE (CBOT)	Futures aggregate global commodity price index
9	COMM	SOYBEANS: #2 YELLOW, 1ST EXPIRATION FUTURE NEARBY - SETTLEMENT - CHICAGO BOARD OF TRADE (CBOT)	Futures aggregate global commodity price index
10	COMM	SILVER BAR: .999 FINE, 1ST EXPIRATION FUTURE NEARBY - SETTLEMENT - COMEX - A DIVISION OF THE NEW YORK MERCANTILE EXCHANGE (NYMEX)	Futures aggregate global commodity price index
11	COMM	GOLD BAR, 0.995 FINE, 1ST EXPIRATION FUTURE NEARBY, SETTLEMENT - COMEX, DIVISION OF NYMEX	Futures aggregate global commodity price index
12	COMM	GAS OIL: 1ST EXPIRATION FUTURE NEARBY - SETTLEMENT - INTERCONTINENTAL EXCHANGE, INC. (ICE)	Futures aggregate global commodity price index
13	COMM	CRUDE OIL	Futures aggregate global commodity price index

14	General	VIX	Chicago Board Options Exchange Volatility Index
15	General	Narrow Money	A category of money supply that includes all physical money like coins and currency along with demand deposits and other liquid assets held by the central bank. In the US narrow money is classified as M1 (M0 + demand accounts), while in the U.K. M0 is referenced as narrow money.
16	General	IRT	Interest Rate
17	General	Term Spread	Term Spread
18	General	Risk Spread	Risk Spread
19	General	Inflation	Inflation Index
20	General	Jan Dummy	Dummy Variable that captures the January effect

Table B.1.: Description and details regarding the data used in chapter 5.

Bibliography

[1] Abadir, K. M. & Rockinger, M. (2003) Density Functionals, with an Option-Pricing Application. *Econometric Theory*, 19 (5), 778-811.

[2] Aït-Sahalia, Y. & Lo, A.W. (1998) Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices. *Joumal of Finance*, 53 (2), 499-547.

[3] Aït-Sahalia, Y. & Lo, A.W. (2000) Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, 94 (1-2), 9-51.

[4] Aït-Sahalia, Y., Mykland, P. A. & Zhang, L. (2005) How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise. *Review of Financial Studies*, 18, 351-416.

[5] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. & Czáki, F. (eds.) *Second International Symposium of Information Theory*, Tsahkadsor, Armenian SSR 1973. pp. 267-281.

[6] Alexander, S. S. (1961) Price Movements in Speculative Markets: Trends or Random Walks. *Industrial Management Review*, 2, 7-26.

[7] Alexander, C. (2001) Orthogonal GARCH. *Mastering Risk*, 2, 21-38.

[8] Alexander, C. & Chibumba, A. (1996) Multivariate orthogonal factor GARCH. *University of Sussex Discussion Papers in Mathematics.*

[9] Amisano, G. (2006) *An Introduction to Bayesian Econometrics for Macroeconomists.* [Lecture notes] A 6 meeting Ph.D. Course, Milan, Bocconi University, January 2006.

[10] Andersen, T. G. & Bollerslev, T. (1998) Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. *International Economic Review*, 39, 885-905.

[11] Andersen, T. G., Bollerslev, T. & Lange, S. (1999) Forecasting Financial Market Volatility: Sample Frequency vis-à-vis Forecast Horizon. *Journal of Empirical Finance*, 6, 457-477.

[12] Andersen, T. G., Bollerslev, T. & Meddahi, N. (2009) Realized Volatility Forecasting and Market Microstructure Noise. *Montreal University working paper*. [Online] Available from: http://www-gremaq.univ-tlse1.fr/perso/meddahi/ABM3_Final.pdf [Accessed 01/11].

[13] Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2000) Great Realizations. *Risk*, 13, 105-108.

[14] Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2001) The Distribution of Exchange Rate Volatility. *Journal of the American Statistical Association*, 96, 42-55.

[15] Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2003) Modeling and Forecasting Realized Volatility. *Econometrica*, 71, 529-626.

[16] Areal, N. M. P. C. & Taylor S. J. (2002) The Realized Volatility of FTSE-100 Futures Prices. *Journal of Futures Market*, 22, 627-648.

[17] Bahra, B. (1996) Probability distributions of future asset prices implied by options prices. *Bank of England Quarterly Bulletin*, 36 (3), 299-311.

[18] Bahra, B. (1997) Implied risk-neutral probability density functions from options prices: Theory and application. *Bank of England working papers*. [Online] 66. Available from: http://www.bankofengland.co.uk/publications/workingpapers/wp66.pdf [Accessed 03/2010]

[19] Baillie, R. T. (1996) Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, 73 (1), 5-59.

[20] Bakshi, G., Kapadia, N. & Madan, D. (2003) Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options. *Review of Financial Studies*, 16 (1), 101-43.

[21] Bali, T. & Weinbaum, D. (2005) Comparative Study of Alternative Extreme Value Volatility Estimators. *Journal of Futures Markets*, 25 (9), 873-892.

[22] Ball, C. & Roma, A. (1994) Stochastic volatility option pricing. *Journal of Financial and Quantitative Analysis*, 29, 589-607.

[23] Bandi, F. & Russell, J. (2006) Separating Microstructure Noise from Volatility. *Journal of Financial Economics*, 79, 655-692.

[24] Bandi, F. & Russell, J. (2008) Microstructure Noise, Realized Volatility, and Optimal Sampling. *Review of Economic Studies*, 75, 339-369.

[25] Bao, Y., Lee, T-H. & Saltoglu, B. (2007) Comparing density forecast models. *Journal of Forecasting*, 26 (3), 203-225.

[26] Barndorff-Nielsen, O. E. & Shephard, N. (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society*, 63 (2), 167–207.

[27] Barndorff-Nielsen, O. E. & Shephard, N. (2002a) Econometric Analysis of Realised Volatility and its Use in Estimating Stochastic Volatility Models. *Journal of the Royal Statistical Society*, 64, 253-280.

[28] Barndorff-Nielsen, O. E. & Shephard, N. (2002b) Estimating quadratic variation using realised variance. *Journal of Applied Econometrics*, 17 (5), 457-477.

[29] Barndorff-Nielsen, O. E. & Shephard, N. (2004a) Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2 (1), 1-37.

[30] Barndorff-Nielsen, O. E. & Shephard, N. (2004b) Econometric analysis of realized covariation: High frequency based covariance, regression and correlation in financial economics. *Econometrica*, 72 (3), 885-925.

[31] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2008) Multivariate Realised Kernels: Consistent Positive Semi-Deffinite Estimators of the Covariation of Equity

Prices with Noise and Non-Synchronous Trading. *Oxford University working paper*, [Online]. Available from: http://www.oxford-man.ox.ac.uk/∼nshephard/papers/kernelMult_25_7_08.pdf [Accessed 01/11].

[32] Barone-Adesi, G., Egle, R. & Mancini, L. (2006) A GARCH Option Pricing Model in Incomplete Markets. *Swiss Finance Institute Research Paper Series, Swiss Finance Institute*. [Online] 07-03. Available from: http://econpapers.repec.org/RePEc:chf:rpseri:rp0703 [Accessed 03/2010].

[33] Basu, S. & Chib, S. (2003) Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models. *Journal of American Statistical Association*, 98 (461), 224-235.

[34] Bates, D. S. (1996a) Dollar jump fears, 1984-1992: Distributional abnormalities implicit in currency futures options. *Journal of International Money and Finance*, 15 (1), 65-93.

[35] Bates, D. S. (1996b) Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *Journal of Financial Studies*, 9 (1), 69-107.

[36] Bauwens, L. & Lubrano, M. (1998) Bayesian Inference on GARCH Models Using the Gibbs Sampler. *CORE Discussion Papers, Universit catholique de Louvain, Center for Operations Research and Econometrics (CORE)*. [Online] 96a21. Available from: http://econpapers.repec.org/RePEc:cor:louvco:1996027 [Accessed 03/2010].

[37] Bauwens, L. & Lubrano, M. (2001) Bayesian option pricing using asymmetric GARCH models. *Journal of Empirical Finance*, 9 (3), 321-342.

[38] Banz, R. & Miller, M. (1978) Prices for State-Contingent Claims: Some Estimates and Applications. *Journal of Business*, 51, 653-672.

[39] Bekaert, G. & Wu, G. (1997) Asymmetric Volatility and Risk in Equity Markets. *The Review of Financial Studies*, 13 (1), 1-42.

[40] Berger, O. J., Pericchi, R. L. (1996) The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, 91 (433), 109-122.

[41] Bingham, N. & Schmidt, R. (2006) *Interplay between Distributional and Temporal Dependence: An Empirical Study with High-frequency Asset Returns*. In: Kabanov, Y., Liptser, R., Stoyanov, J. (eds), From Stochastic Calculus to Mathematical Finance, pp. 69-90. Germany, Springer Berlin Heidelberg.

[42] Black, F. & Litterman, R. (1992) Global Portfolio Optimization. *Financial Analyst Journal*, 48 (5), 28-43.

[43] Black, F. & Scholes, M. (1973) The pricing of options and corporate liabilities. *Journal of Political Economy*, 87, 637-659.

[44] Blair, B. J., Poon, S. & Taylor, S. J. (2001) Modelling S&P 100 volatility : the information content of stock returns. *Journal of Banking and Finance*, 25, 1665-1679.

[45] Bliss, R. & Panigirtzoglou, N. (2004) Option Implied Risk Aversion Estimates. *Journal of Finance*, 59 (1), 407-446.

[46] Bollerslev, T., (1986) Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* , 31 (3), 307-327.

[47] Bollerslev, T., Mikkelsen, H. O. (1996) Modelling and pricing long memory in stock market volatility. *Journal of Econometrics*, 73 (1), 151-184.

[48] Bondarenko, O. (2003) Estimation of risk-neutral densities using positive convolution approximation. *Journal of Econometrics*, 116 (1-2), 85-112.

[49] Bookstaber, R. & Mcdonanld, J. (1987) General Distribution for Describing Security Price Returns. *Journal of Business*, 60 (3), 401-424.

[50] Breeden, D. & Litzenberger, R. (1978) Prices of State Contingent Claims Implicit in Options Prices. *Journal of Business*, 51 (4), 621-651.

[51] Brigo, D., Dalessandro, A., Neugebauer, M. & Triki, F. (2008) A Stochastic Processes Toolkit for Risk Management. *Quantitative Finance Papers*. [Online] 0812. Available from: http://arxiv.org/PS_cache/arxiv/pdf/0812/0812.4210v1.pdf [Accessed 01/2010].

[52] Britten-Jones, M. & Neuberger, A. (2000) Option Prices, Implied Price Processes, and Stochastic Volatility. *Journal of Finance*, 55, 839-866.

[53] Brooks, S. P. & Gelman, A. (1998) General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational Finance*, 7 (4), 434-455.

[54] Brooks, C. & Henry, O. T. (2000) Can portmanteau nonlinearity tests serve as general mis-specification tests?: Evidence from symmetric and asymmetric GARCH models. *Economics Letters*, 67 (3), 245-251.

[55] Buchen, P. & Kelly, M. (1996) The Maximum Entropy Distribution of an Asset Inferred from Option Prices. *Journal of Financial and Quantitative Analysis*, 31 (1), 143-159.

[56] Campbell, Y. J. & Shiller, J. R. (1991) Yield Spreads and Interest Rate Movements: A Bird's Eye View. *Review of Economic Studies*, 58 (3), 495-514.

[57] Campa, J. M., Chang, P. H. K. & Refalo, J. F. (1999) An Options-Based Analysis of Emerging Market Exchange Rate Expectations: Brazil's Real Plan, 1994-1997. *The National Bureau of Economic Research*. [Online] 6929. Available from: http://www.nber.org/papers/w6929 [Accessed 03/2010].

[58] Candes, E. & Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35 (6), 2313-2351.

[59] Canian, L. & Figlewski, S. (1993) The information content of implied volatility. *Review of Financial Studies*, 6, 659-681.

[60] Cappiello, L., Engle, F. R. & Sheppard, K. (2006) Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns. *Journal of Financial Econometrics*, 4 (4), 537-572.

[61] Carhart, M. M. (1997) On persistence in mutual fund performance. *Journal of Finance*, 52, 57-82.

[62] Carlin, P. B., Louis, A. T., Carlin B. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*. New York, Chapman & Hall.

[63] Chen, Y. (2004) Exchange Rates and Fundamentals: Evidence from Commodity Economies. *Mimeograph, Harvard University*. [Online] Available from: http://faculty.washington.edu/yuchin/Papers/ner.pdf [Accessed 03/2010].

[64] Chen, Y. & Rogoff, K. (2003) Commodity Currencies. *Journal of International Economics*, 60 (1), 133-160.

[65] Chen, Y. & Rogoff K. (2006) Are Commodity Currencies an Exception to the Rule? *Mimeograph, Harvard University*. [Online] Available from: http://faculty.washington.edu/yuchin/Papers/Chen-Rogoff.pdf [Accessed 03/2010].

[66] Chen, Y., Rogoff, K. and Rossi, B. (2008) Can Exchange Rates Forecast Commodity Prices? *The National Bureau of Economic Reserach*. [Online] 13901. Available from: http://www.nber.org/papers/w13901 [Accessed 03/2010].

[67] Chib, S. & Greenberg, E. (1995) Understanding the Metropolis-Hastings Algorithm. *Journal of the American Statistical Association*, 49 (4), 327-335.

[68] Chib, S. & Jelianzkov, I. (2001) Marginal Likelihood from the Metropolis-Hastings Output. *Journal of American Statistical Association*, 96 (453), 270-281.

[69] Christoffersen, P. & Jacobs, K. (2004) Which GARCH Model for Option Valuation? *Management Science*, 50 (9), 1204-1221.

[70] Christoffersen, P. & Mazzotta, S. (2005) The accuracy of density forecasts from foreign exchange options. *Journal of Financial Econometrics*, 3 (4), 578-605.

[71] Christoffersen, P., Heston, S. & Jacobs, K. (2006) Option Valuation with Conditional Skewness. *Journal of Econometrics*, 131 (1-2), 253-284.

[72] Concepción Ausín, M. & Galeano, P. Bayesian estimation of the Gaussian mixture GARCH model. *Statistics and Econometrics Working Papers, Universidad Carlos III, Departamento de Estadstica y Econometra*, [Online] ws053605. Available from: http://econpapers.repec.org/RePEc:cte:wsrepe:ws053605 [Accessed 03/2010].

[73] Corradi V. (2000) Reconsidering the continuous time limit of the GARCH(1,1) process. *Journal of Econometrics*, 96,145-153.

[74] Corrado, C. J. (2001) Option Pricing based on the Generalized Lambda Distribution. *Journal of Futures Markets*, 21 (3), 213-236.

[75] Corrado, C. J. & Su, T. (1997) Implied volatility skews and stock index skewness and kurtosis implied by S&P 500 Index option prices. *Journal of Derivatives*, 4 (4), 8-19.

[76] Cowles, M. K. & Carlin, B. P. (1996) Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91 (434), 883-904.

[77] Daniel, K., Hirshleifer, D. & Subrahmanyam, A. (1998) Investor Psychology and Security Market Under and Overreactions. *Journal of Finance*, 53 (6), 1839-1885.

[78] de Goeij, P. & Marquering, W. (2005) The generalized asymmetric dynamic covariance model. *Finance Research Letters*, 2 (2), 67-74.

[79] Deo, R., Hurvich, C. & Lu, Y. (2006) Forecasting realized volatility using a long-memory stochastic volatility model: estimation, prediction and seasonal adjustment. *Journal of Econometrics*, 131 (1-2), 29-58.

[80] Diciccio, J. T., Kass E. R., Raftery, A. & Wasserman, L. (1997) Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association*, 92 (439), 903-915.

[81] Diebold, F. X., Hahn, J. & Tay, A. (1999) Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns on Foreign Exchange. *Review of Economics and Statistics*, 81 (4), 661-673.

[82] Diebold, F. X., Tay, A. & Wallis, K. (1999) Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters. *National Bureau of Economic Research, Inc.* [Online] 6228. Available from: http://www.nber.org/papers/w6228.pdf [Accessed 01/2010].

[83] Diebold, F. X., Gunther, T. & Tay, A. (1998) Evaluating Density Forecasts, with Applications to Financial Risk Management. *International Economic Review*, 39 (4), 863-883.

[84] Ding, Z., Granger, W. J. C. & Engle, F. R. (1993) A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1 (1), 83-106

[85] Ding, Z. & Granger, C. W. J. (1996) Modelling volatility persistence of speculative returns: A new approach. *Journal of Econometrics*, 73 (1), 185-215.

[86] Dotsis, G. & Markellos, R. (2007) The Finite Sample Properties of GARCH Option Pricing Model. *Journal of Future Markets*, 27 (6), 599-615.

[87] Draper, D. (1995) Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society*, 57 (1), 45-97.

[88] Draper, N. R. & Smith, H. (1998) *Applied regression analysis.* 3rd edition, New York, John Wiley & Sons, Inc.

[89] Duan, J. (1995) The GARCH option pricing model. *Mathematical Finance*, 5 (1), 13-32.

[90] Duan, J. (1997) Augmented GARCH(p,q) process and its diffusion limit. *Journal of Econometrics*, 79 (1), 97-127.

[91] Duan, J., Gauthier, G. & Simonato, J. (1999) An analytical approximation for the GARCH option pricing model. *Journal of Computational Finance*, 2 (4), 75-116.

[92] Duan, J., Gauthier, G., Sasseville, C. & Simonato, J. (2005) Approximating the GJR-GARCH and EGARCH Option Pricing Models Analytically. *Journal of Computational Finance*, 9 (3), 41-70.

[93] Dumas, B., Fleming, J. & Whaley, R. E. (1998) Implied Volatility Functions: Empirical Tests. *The Journal of Finance*, 53 (6), 2059-2106.

[94] Dušan I., D. & Pérignon, C. (2000) Evolution of Market Uncertainty around Earnings Announcements. *Journal of Banking & Finance*, 25 (9), 1769-1788.

[95] Efron, B. & Tibshirani, R. (1993) *An Introduction to the Bootstrap*. London, Chapman & Hall/CRC.

[96] Efron, B., Hastie, T., Johnstone, I. & Tibshirani R. (2004) Least angle regression. *Annals of Statistics*, 32, 407-451.

[97] Engle, F. R. (1982) Autoregressive Conditional Heteroskedasticity with Estimates of the variance of U.K. Inflation. *Econometrica*, 50 (4), 987-1008.

[98] Engle, F. R. & Ng, K. V. (1993) Measuring and Testing the Impact of News on Volatility. *The Journal of Finance*, 48 (5), 1749-1778.

[99] Engel, C. & West, D. K. (2004) Exchange Rates and Fundamentals. *National Bureau of Economic Research*. [Online] 10723. Available from: http://www.nber.org/papers/w10723.pdf [Accessed 03/2010].

[100] Fama, F. E. (1970) Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25 (2), 383-417.

[101] Fama, F. E. & French, R. K. (1988) Dividend yields and expected stock returns. *Journal of Financial Economics*, 22 (1), 3-25.

[102] Fama, E. F. (1991) Efficient capital markets: II. *Journal of Finance*, 46 (5), 1575-1617.

[103] Fama, E. F. & French, K. R. (1993) Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33 (1), 3-56.

[104] Fama, E. F. & French, K. R. (1992) The Cross-Section of Expected Stock Returns, *Journal of Finance*, 47 (2), 427-465.

[105] Figlewski, S. (2009) Estimating the Implied risk-neutral Density for the U.S. Market Portfolio. Forthcoming in *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle.*

[106] Fleming, J., Kirby, C. & Ostdiek, B. (2003) The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics*, 67, 473-509.

[107] Fouque, J., Papanicolaou, G. & Sircar, K. (1998) Asymptotics of a two-scale stochastic volatility model. *Equations aux derivees partielles et applications in hounour of Jacques-Louis Lions, May*, 517526.

[108] Fournie, E., Lebuchoux, J. & Touzi, N. (1997). Small noise expansion and importance sampling. *Asymptotic Analysis*, 14, 361376.

[109] Fraley, C. & Hesterberg, T. (2009) Least-angle regression and Lasso for large datasets. *Statistical Analysis and Data Mining archive*, 1 (4), 251-259.

[110] Gander, M. P. S. (2004) *Inference for Stochastic Volatility Models Driven by Lévy Processes.* A thesis submitted for the degree of Doctor of Philosophy of the University of London and for the Diploma of Imperial College, London, Imperial College.

[111] Gatheral, J. (2006) *The Volatility Surface: A Practitioner's Guide.* USA, John Wiley & Sons, Ltd.

[112] Gatheral, J. (2008) *Consistent Modelling of SPX and VIX Options.* [Presentation] The Fifth World Congress of the Bachelier Finance Society, London, 18th July.

[113] Gelfand, A. E. & Dey, D. (1994) Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Sociey Series B*, 56 (3), 501-514.

[114] Geweke, J. (1989) Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica*, 57 (6), 1317-1339.

[115] Geweke, J. (1997) Posterior Simulators in Econometrics. *Advances in economics and econometrics: Theory and applications*, 3, 128-165.

[116] Geweke, J. (1999) Using simulation methods for Bayesian econometric models: inference, development, and communication. *Econometric Reviews*, 18 (1), 1-73.

[117] Geweke, J. (2000) Simulation-Based Bayesian Inference for Economic Time Series. *Federal Reserve Bank of Minneapolis*. [Online] 570, 255-299. Available from: http://www.minneapolisfed.org/research/WP/WP570.pdf [Accessed 03/2010].

[118] Geweke, J. (2005) *Contemporary Bayesian Econometrics and Statistics*. Hoboken, John Wiley & Sons, Inc.

[119] Ghosh, S. (2007) Adaptive elastic net: An improvement of elastic net to achieve oracle properties. *Department of Mathematical Sciences, Indiana University-Purdue University, Indianapolis*. [Online]. Available from: http://www.math.iupui.edu/ samiran/Preprint/adelnet.pdf [Accessed 01/2010].

[120] Giacomini, R. (2002) Comparing Density Forecasts via Weighted Likelihood Ratio Tests: Asymptotic and Bootstrap Methods. *Boston College Working Papers in Economics*. [Online] 583. Available from: http://fmwww.bc.edu/EC-P/WP583.pdf [Accessed 01/2010].

[121] Giacomini, R. & Komunjer, I. (2005) Evaluation and Combination of Conditional Quantile Forecasts. *Journal of Business and Economic Statistics*, 23 (4), 416-431.

[122] Giamouridis, D. (2005) Inferring option-implied investors' risk preferences. *Applied Financial Economics*, 15 (7), 479-488.

[123] Giamouridis, D. & Tamvakis, M. (2002) Asymptotic Distribution Expansions in Option Pricing. *Journal of Derivatives*, 9 (4), 33-44.

[124] Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. London, Chapman & Hall.

[125] Giot, P. & Laurent, S. (2007) The Information Content of Implied Volatility in Light of the Jump/Continuous Decomposition of Realized Volatility. *Journal of Futures Markets*, 27, 337-359.

[126] Glosten, L., Jaganathan, R. & Runkle, D. (1993) On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance*, 48 (5), 1779-1801.

[127] Granger, C. W. J. & Jeon, Y. (2004) Thick modelling. *Economic Modelling*, 21 (2), 323-343.

[128] Gray, S. (1996) Modelling the Conditional Distribution of Interest Rates as a Regime-Switching Process. *Journal of Financial Economics*, 42 (1), 27-62.

[129] Griffin, J. E. & Steel, M. F. J. (2002) Inference with non-Gaussian Ornstein-Uhlenbeck processes for stochastic volatility. *EconWPA*. [Online] 0201002. Available from: http://129.3.20.41/eps/em/papers/0201/0201002.pdf [Accessed 01/2010].

[130] Grünbichler, A. & Longstaff, F. A. (1995) Valuing Futures and Options on Volatility. *Journal of Banking and Finance*, 20 (6), 985-1001.

[131] Hansen, E. B. (1994) Autoregressive Conditional Density Estimation. *International Economic Review*, 35 (3), 705-730.

[132] Hamilton, J. & Susmel, R. (1994) Autoregressive Conditional Heteroscedasticity and changes in Regime. *Journal of Econometrics*, 64 (1-2), 307-333.

[133] Hastie, T., Taylor, J., Tibshirani, R. & Walther, G. (2007) Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*. [Online] 1, 1-29. Available from: www.i-journals.org/ejs/include/getdoc.php?id=74 [Accessed 01/2010].

[134] Hawkins, R. (1997) Maximum entropy and derivative securities. *Advances in Econometrics*, 12, 277-301.

[135] Hentschel, L. (1995) All in the family Nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics*, 39 (1), 71-104.

[136] Hesterberg, T., Choi, N., Meier, L. & Fraley, C. (2008) Least angle and ℓ_1 regression: A Review. *Statistics Surveys*, 2, 61-93.

[137] Heston, S. (1993) A closed-form solution for option with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6 (2), 327-343.

[138] Heston, S. & Nandi, S. (2000) A closed-form GARCH option valuation model. *Review of Financial Studies*, 13 (3), 585-625.

[139] Higgins, M. & Bera, A. (1992) A Class of Nonlinear ARCH Models. *International Economic Reviews*, 33 (1), 137-158.

[140] Hills, S. E. & Smith, A. F. M. (1992), Parametrisation Issues in Bayesian Inference. *Journal of Bayesian Statistics*, 4, 227-246.

[141] Hirshleifer, D. (2001) Investor Psychology and Asset Pricing. *Journal of Finance*, 56 (4), 1533-1597.

[142] Hoeting, A. J., Madigan, D., Raftery, E. A. & Volinsky, T. C. (1999) Bayesian Model Averaging: A Tutorial. *Journal of Statistical Science*, 14 (4), 382-417.

[143] Hsieh, K. C. & Ritchken, P. (2006) An Empirical Comparison of GARCH Option Pricing Models. *Review of Derivatives Research*, 8 (3), 129-150.

[144] Hull, J. & White, A. (1987) The pricing of options on assets with stochastic volatilities. *Journal of Finance*, 42 (2), 281-300.

[145] Ishwaran, H. (2004) Discussion of "Least Angle Regression" by Efron *et al. Annals of Statistics*, 32 (2), 452-458.

[146] Jackwerth, J. C. (1996) Generalised binomial trees. *Journal of Derivatives*, 5 (2), 7-17.

[147] Jackwerth, J. C. (1999) Option Implied Risk-Neutral Distributions and Implied Binomial Trees: A Literature Review. *Journal of Derivatives*, 7 (2), 66-82.

[148] Jackwerth, J. C. & Rubinstein, M. (1996) Recovering probability distributions from option prices. *Journal of Finance*, 51 (5), 1611-1631.

[149] Jackwerth, J. C. & Carsten, J. (2000) Recovering Risk Aversion from Option Prices and Realized Returns. *Review of Financial Studies*, 13 (2), 433-51.

[150] Jarque, C. M. & Bera, A. K. (1987) A test for normality of observations and regression residuals. *International Statistical Review*, 55 (2), 1-10.

[151] Jarrow, R. & Rudd, A. (1982) Approximate Valuation for Arbitrary Stochastic Processes. *Journal of Financial Economics*, 10 (3), 347-369.

[152] Jaynes, E. T. (1957) Information theory and statistical mechanics. *Physics Review*, 106, 620-630.

[153] Jeffreys, H. (1961) Theory of Probability. 3rd edition, New York, Oxford University Press.

[154] Jiang, G. J. & Tian, Y. S. (2005) The Model-Free Implied Volatility and Its Information Content. *Review of Financial Studies*, 18 (4), 1305-1342.

[155] Kass, E. R. & Raftery, E. A. (1993) Bayes factors. *Journal of the American Statistical Association*, 90 (490), 773-795.

[156] Kass, E. R. & Wasserman, L. (1995) A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, 90 (431), 928-934.

[157] Kawakatsu, H. (2006) Matrix exponential GARCH. *Journal of Econometrics*, 134 (1), 95-128.

[158] Kendall, M. G. (1953) The Analysis of Economic Time-Series-Part I: Prices. *Journal of the Royal Statistical Society*, 116 (1), 11-34.

[159] Kilic, E. (2005) Forecasting Volatility of Turkish Markets: A Comparison of Thin and Thick Models. *EconWPA*. [Online] 0510007. Available from: http://129.3.20.41/eps/em/papers/0510/0510007.pdf [Accessed 03/2010].

[160] Klüppelberg C., Lindner, A. & Maller, R. (2004) A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour. *Journal of Applied Probability*, 41 (3), 601-622.

[161] Klüppelberg, C., Lindner, A. & Maller, R. (2004) *Continuous time volatility modelling: COGARCH versus Ornstein-Uhlenbeck models.* In: Kabanov, Y., Lipster, R. and Stoyanov, J. (eds.) From Stochastic Calculus to Mathematical Finance: the Shiryaev Festschrift, pp. 393-419. Berlin, Springer.

[162] Knight, K. (2004) Discussion of "Least Angle Regression" by Efron *et al. Annals of Statistics*, 32 (2), 458-460.

[163] Koopman, S. J., Jungbacker, B. & Hol, E. (2005) Forecasting daily variability of the S&P100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance*, 12 (3), 445-475.

[164] Kroner, K. F. & Ng, V. K. (1998) Modelling asymmetric comovements of asset returns. *Review of Financial Studies*, 11 (4), 817844.

[165] Lagarias, J. C., Reeds, J. A., Wright, M. H. & Wright, P. E. (1998) Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, 9 (1), 112-147.

[166] Lee, S., Lee, H., Abeel, P. & Ng. A. (2006) Efficient $\ell 1$ regularized logistic regression. Proceedings of the *Twenty-First National Conference on Artificial Intelligence (AAAI-06).*

[167] Li, C. (2009) Efficient Valuation of VIX Options under Gatheral's Double Mean-Reverting Stochastic Volatility Model. *SSRN*. [Online] 1416608. Available from: http://ssrn.com/abstract=1416608 [Accessed 01/2010].

[168] Lindner, A. M. (2009) *Continuous time approximations to GARCH and stochastic volatility models.* In: Andersen, T. G., Davis, R. A.,

Krei, J.-P. and Mikosch, T. (eds.), Handbook of Financial Time Series, pp. 481-496. Berlin, Springer.

[169] Liu, X., Shackleton, M., Taylor, S. & Xu, X. (2007) Closed-form transformations from risk-neutral to real-world distributions. *Journal of Banking and Finance*, 31 (5), 1501-1520.

[170] Lo, W. A. (2004) The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective. *Journal of Portfolio Management*, 30, 15-29.

[171] Lo, W. A. (2005) Reconciling Efficient Markets with Behavioural Finance: The Adaptive Markets Hypothesis. *Journal of Investment Consulting*, 7, 21-44.

[172] Lo, W. A. & MacKinlay A. C. (1999) *A Non-Random Walk Down Wall Street*. Princeton, Princeton University Press.

[173] Lo, W. A., Mamaysky, H. & Wang, J. (2000) Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. *Journal of Finance*, 55, 1705-1765.

[174] Loubes, J. & Massart, P. (2004) Discussion of "least angle regression" by Efron *et al. Annals of Statistics*, 32, 460-465.

[175] Longstaff, F. (1995) Option pricing and the martingale restriction. *Review of Financial Studies*, 8 (4), 1091-1124.

[176] Madigan, D. & Raftery, A. E. (1994) Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window. *Journal of the American Statistics Association*, 89, 1535-1546.

[177] Madigan, D. & York, J. (1995) Bayesian Graphical Models for Discrete Data. *International Statistical Review*, 63 (2), 215-232.

[178] Madigan, D. & Ridgeway, G. (2004) Discussion of Least Angle Regression by Efron *et al. Annals of Statistics*, 32 (2), 465469.

[179] Malz, A. M. (1997a) Estimating the Probability Distribution of the Futures Exchange Rate from Option Prices. *Journal of Derivatives*, 5 (2), 18-36.

[180] Malz, A. M. (1997b) *Option-based Estimates of the Probability Distribution of Exchange Rates and Currency Excess Returns*. Federal Reserve of New York. Report number: 32.

[181] Mandelbrot, B. (1965) Forecasts of Future Prices, Unbiased Markets, and Martingale Models. *Journal of Business*, 39, 242-255.

[182] Markose, S. & Alentorn, A. (2005) The Generalized Extreme Value (GEV) Distribution, Implied Tail Index and Option Pricing. *Economics Discussion Papers, University of Essex, Department of Economics*. [Online] 594. Available from: http://ideas.repec.org/p/esx/essedp/594.html [Accessed 01/2010].

[183] Marquardt, D. (1970) Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, 12 (3), 591-612.

[184] Martens, M. (2002) Measuring and Forecasting S&P500 Index Futures Volatility using High-Frequency Data. *Journal of Futures Markets*, 22, 497-518.

[185] Martens, M. & Zein, J. (2004) Predicting financial volatility: High-frequency time-series forecasts vis-à-vis implied volatility. *Journal of Futures Markets*, 24 (11) 1005-1028.

[186] Martin, G. M., Reidy, A. & Wright J. (2009) Does the Option Market Produce Superior Forecasts of Noise-Corrected Volatility Measures? *Journal of Applied Econometrics*, 24, 77-104.

[187] Massey, F. J. (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46 (253), 68-78.

[188] Masters, T. (1993) *Practical Neural network Recipes in C++*. San Diego, Academic Press Professional, Inc.

[189] McDonald, J. B. & Xu, Y. J. (1995) A generalization of the beta distribution with applications. *Journal of Econometrics*, 69 (2), 427-428.

[190] Meddahi, N. (2002) A Theoretical Comparison Between Integrated and Realized Volatility. *Journal of Applied Econometrics*, 17, 479-508.

[191] Meier, L. & Bühlmann, P. (2007) Smoothing ℓ1-penalized estimators for high-dimensional time-course data. *Electronic Journal of Statistics*. [Online] 1, 597-615. Available from: http://www.i-journals.org/ejs/include/getdoc.php?id=520&article=103 [Accessed 03/2010].

[192] Meinshausen, N. (2007) LASSO with Relaxation. *Computational Statistics & Data Analysis*, 52 (1), 374-393.

[193] Meinshausen, N., Rocha, G. & Yu, B. (2007) A tale of three cousins: Lasso, ℓ_2 Boosting, and Dantzig. *Annals of Statistics*, 35 (6), 2373-2384.

[194] Melick, W. R. & Thomas, C. P. (1997) Recovering an asset's implied PDF from options prices: An application to crude oil during the Gulf crisis. *Journal of Financial and Quantitative Analysis*, 32 (1), 91-116.

[195] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemistry Physics*, 21 (6), 1087-1092.

[196] Mira, A. (2004) *Introduction to Monte Carlo and MCMC Methods*. [Lecture notes] Bayesian methods in Inverse Problems Summer course, Kuopio, Finland, 21-24/6/2004.

[197] Mukhopadhyay, N., Ghosh, K. J. & O. Berger, J. (2005) Some Bayesian Predictive Approaches to Model Selection. *Statistics & Probability Letters Statistics*, 73 (4), 369-379.

[198] Nakatuma, T. & Tsurumi, H. (undated) ARMA-GARCH Models: Bayes Estimation versus MLE, and Bayes Non-Stationarityf Test. *Departmental Working Papers, Rutgers University, Department of Economics*. [Online] 199619. Available from: http://econpapers.repec.org/RePEc:rut:rutres:199619 [Accessed 03/2010].

[199] Nakatuma, T. (1998) A Markov-Chain Sampling Algorithm for GARCH Models. *Studies in Nonlinear Dynamics and Econometrics*, 3 (2), 107-117.

[200] Nakatuma, T. (2000) Bayesian Analysis of ARMA-GARCH models: A Markov Chain Sampling Approach. *Journal of Econometrics*, 95 (1), 57-69.

[201] Neely, C. J. (2002) Forecasting foreign exchange volatility: Is implied volatility the best that we can do?. *Federal Reserve Bank of St. Louis working paper*. [Online] 2002-017D. Available from: http://research.stlouisfed.org/wp/2002/2002-017.pdf [Accessed 01/2011].

[202] Nelson, D. B. (1991) Conditional heteroskedasticity in asset returns: A New Approach. *Econometrica*, 59 (2), 347-370.

[203] Odaki, M. (1993) On the invertibility of fractionally differenced ARIMA processes. *Biometrika*, 80(3), 703709.

[204] Odean, T. (1999) Do Investors Trade Too Much? *American Economic Review*, 89, 1279-1298.

[205] O'Hagan, A. (1995) Fractional Bayes Factors for Model Comparion. *Journal of the Royal Statistical Society*, 57 (1), 99-138.

[206] Osborne, M. F. M. (1959) Brownian Motion in the Stock Market. *Operations Research*, 7 (2), 145-173.

[207] Patton, A. (2006) *Forecasting Financial Time Series*. [Lecture notes] Time Series Analysis, London School of Economics and Political Science, January - March 2006.

[208] Patton, A. (2010) *Andrew Patton's Matlab code page*. [Online] Available from: http://econ.duke.edu/~ap172/code.html [Accessed 03/2010].

[209] Park, M. Y. & Hastie, T. (2007) An $\ell 1$ regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B*, 69, 659-677.

[210] Pesaran, M. H. & Timmermann A. (1995) Predictability of Stock Returns: Robustness and Economic Significance. *The Journal of Finance*, 50 (4), 1201-1228.

[211] Pesaran, M. H. & Timmermann, A. (2002) Market Timing and Return Prediction under Model Instability. *Journal of Empirical Finance*, 9 (5), 495-510.

[212] Pesaran, M. H. & Zaffaroni, P. (2004) Model Averaging and Value-at-Risk Based Evaluation of Large Multi Asset Volatility Models for Risk Management. *CESifo Working Paper Series, CESifo Group Munich.* [Online] 1358. Available from: http://econpapers.repec.org/RePEc:ces:ceswps:_1358 [Accessed 03/2010].

[213] Pesaran, M. H. & Zaffaroni, P. (2008) Optimal Asset Allocation with Factor Models for Large Portfolios. *Cambridge Working Papers in Economics.* [Online] 0813. Available from: http://www.econ.cam.ac.uk/dae/repec/cam/pdf/cwpe0813.pdf [Accessed 03/2010].

[214] Pong, S., Shackleton, M. B., Taylor, S. J. & Xu, X. (2004) Forecasting currency volatility: a comparison of implied volatilities and AR(FI)MA models. *Journal of Banking and Finance*, 28, 2541-2563.

[215] Rabemananjara R. & Zakoian, J. M. (1993) Threshold Arch Models and Asymmetries in Volatility. *Journal of Applied Econometrics*, 8 (1), 31-49.

[216] Raftery, E. A., Madigan, D. & Hoeting, A. J. (1993) *Model Selection and Accounting for Model Uncertainty in Linear Regression Models.* Department of Statistics, GN-22 University of Washington Seattle, Washington 98195 USA. Report number: 262.

[217] Raftery, E. A., Madigan, D. & Hoeting, A. J. (1997) Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92 (437), 179-191.

[218] Raftery, E. A (1999) Bayes Factors and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection". *Sociological Methods & Research*, 27 (3), 411-427.

[219] Ritchey, R. J. (1990) Call Option Valuation for Discrete Normal Mixtures. *Journal of Financial Research*, 13 (4), 285-296.

[220] Roberts, H. (1959) Stock-market patterns and financial analysis: methodological suggestions. *Journal of Finance*, 14 (1), 1-10.

[221] Rosenberg, J. V. & Engle R. F. (2002), Empirical pricing kernels. *Journal of Financial Economics*, 64 (3), 341-372.

[222] Rosenblatt, M. (1952) Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23 (3), 470-472.

[223] Ross, S. (1976) Options and Efficiency. *Quarterly Journal of Economics*, 90 (1), 75-89.

[224] Rosset, S. & Zhu, J. (2004) Discussion of "Least Angle Regression" by Efron *et al. Annals of Statistics*, 32, 469-475.

[225] Rubinstein, M. (1983) Displaced Diffusion Option Pricing. *Journal of Finance*, 38 (1), 213-217.

[226] Rubinstein, M. (1994) Implied binomial trees. *Journal of Finance*, 49 (3), 771-818.

[227] Rubinstein, M. (1998) Edgeworth Binomial Trees. *Journal of Derivatives*, 5 (3), 20-27.

[228] Samuelson, P. A. (1965) Proof That Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review*, 6 (2), 41-49.

[229] San Martini, A. & Spezzaferri, F. (1984) A Predictive Model Selection Criterion. *Journal of the Royal Statistical Society*, 46 (2), 296-303.

[230] Savickas, R. (2004) Evidence on Delta Hedging and Implied Volatilities for the Black-Scholes, Gamma and Weibull Option-Pricing Models. *Journal of Financial Research*, 28 (2), 299-317.

[231] Schmidt, M., Fung, G. & Rosales, R. (2007) Fast optimisation methods for ℓ_1 regularization. In the Proceedings of the 18th *European Conference on Machine Learning*. [Online] 4701, 286-297. Available from: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.7676 [Accessed 03/2010].

[232] Schmidt, M. W., Niculescu-Mizil, A. & Murphy, K. P. (2007) Learning graphical model structure using $\ell 1$ regularization paths.

Proceedings of the *Association for the Advancement of Artificial Intelligence AAAI*. [Online] 202, 1278-1283. Available from: https://www.aaai.org/Papers/AAAI/2007/AAAI07-202.pdf [Accessed 03/2010].

[233] Schwarz, G. (1978) Estimating the dimension of a model. *Journal of Annual Statistics*, 6 (2), 461-464.

[234] Ser-Huang Poon (2005) *A Practical Guide to Forecasting Financial Market Volatility*. West Sussex, John Wiley & Sons, Ltd.

[235] Shackleton M. B., Taylor S. J. & Yu P. (2010) A multi-horizon comparison of density forecasts for the S&P 500 using index returns and option prices. *Journal of Banking and Finance*, 34 (11), 2678-2693.

[236] Sheppard, K. (2010) *MFE Financial Econometrics*. [Online]. Available from: http://www.kevinsheppard.net/wiki/Category:MFE [Accessed 03/2010].

[237] Sherrick, B. J., Garcia, P. & Tirupattur, V. (1996) Recovering Probabilistic Information from Option Markets: Tests of Distributional Assumptions. *Journal of Futures Markets*, 16 (5), 545-560.

[238] Shiller, R. J. (1981) Do Stock Prices Move So Much to Be Justified by Subsequent Changes in Dividends? *American Economic Review*, 71 (3), 421-36.

[239] Shiller, R. J. (2000) *Irrational Exuberance*. 2nd edition. Princeton, Princeton University Press.

[240] Shleifer, A. (2000) *Inefficient Markets: An Introduction to Behavioral Finance*. New York, Oxford University Press.

[241] Shimko, D. (1993) Bounds of probability. *Risk Magazine*, 6 (4), 33-37.

[242] Smith, A. F. M. & Spiegelhalter, D. J. (1980) Bayes Factors and Choice Criteria for Linear Models. *Journal of the Royal Statistical Society*, 42 (2), 213-220.

[243] Smith, A. F. M. & Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods.

Journal of Royal Statistical Society, Series B (Methodological), 55 (1), 3-24.

[244] Söderlind, P. & Svensson, L. E. O. (1997) New techniques to extract market expectations from financial instruments. *Journal of Monetary Economics*, 40 (2), 383-429.

[245] Stavarek, D. (2004) Linkages between Stock Prices and Exchange Rates in the EU and the United States. *EconWPA*. [Online] 0406006. Available from: http://129.3.20.41/eps/fin/papers/0406/0406006.pdf [Accessed 03/2010].

[246] Stein, E. & Stein, J. (1991) Stock price distributions with stochastic volatility: An analytic approach. *The Review of Financial Studies*, 4 (4), 727-752.

[247] Stine, R. A. (2004) Discussion of "Least Angle Regression" by Efron *et al.*". *Annals of Statistics*, 32, 475-481.

[248] Stone, M. (1979) Comments on Model Selection Criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society, Series B (Methodological)*, 41 (2), 276-278.

[249] Stutzer, M. (1996) A simple nonparametric approach to derivative security valuation. *Journal of Finance*, 51 (5), 1633-1652.

[250] Taylor, S. J. (2008) *Modelling Financial Time Series*. 2nd edition, Singapure, World Scientific Publishing Co. Pte. Ltd.

[251] Taylor, S. J. & Xu, X. (1997) The incremental volatility information in one million foreign exchange quotations. *Journal of Empirical Finance*, 4, 317-340.

[252] Teräsvirta T. (2009) *An introduction to univariate GARCH models*. In: Torben G. Andersen, Richard A. Davis, Jens-Peter Kreiss, and Thomas Mikosch (eds.) Handbook of Financial Time Series. New York, Springer, 17-42.

[253] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58 (1), 267-288.

[254] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society*, 67 (1), 91-108.

[255] Tierney, L. & Kadane, J. B. (1986) Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81 (393), 82-86.

[256] Tse, Y. K. & Tsui, A. K. C, (2002) A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model with Time-Varying Correlations. *Journal of Business & Economic Statistics*, 20 (3), 351-62.

[257] van der Weide, R. (2002) Generalized Orthogonal GARCH. A Multivariate GARCH model. *Journal of Applied Econometrics*, 17 (5), 549-564.

[258] Volinsky, C. (2010). *Bayesian Model Averaging Home Page*. [Online] Available from: http://www2.research.att.com/ volinsky/bma.html [Accessed 03/2010].

[259] Volinsky, C., Madigan, D., Raftery, E. A. & Kronmal, A. R. (1997) Bayesian Model Averaging in Proportional Hazard Models: Predicting the Risk of a Stroke. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 46 (4), 443-448.

[260] Vrontos, I. D., Delaportas, P. & Politis, D. N. (2000) Full Bayesian Inference for GARCH and EGARCH models. *Journal of Business and Economic Statistics*, 18 (2), 187-198.

[261] Vrontos, I. D., Dellaportas, P. & Politis, D. N. (2003) A full-factor multivariate GARCH model. *Econometrics Journal, Royal Economic Society*, 6 (2), 312-334.

[262] Wasserman, L. (2000) Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, 44 (1), 92-107.

[263] Wiggins, J. (1987). Option values under stochastic volatility. *Journal of Financial Economics*, 19, 351372.

[264] Yuan, M. & Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68 (1), 49-68.

[265] Zhang, L. (2006) Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach. *Bernoulli*, 12, 1019-1043.

[266] Zhang, L., Mykland, P. A. & Aït-Sahalia, Y. (2005) A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data. *Journal of the American Statistical Association*, 100, 1394-1411.

[267] Zhou, B. (1996) High-Frequency Data and Volatility in Foreign Exchange Rates. *Journal of Business and Economic Statistics*, 14, 45-52.

[268] Zou, H. (2006) The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.

[269] Zou, H. & Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67 (2), 301-320.

[270] Zou, H., Hastie, T. & Tibshirani, R. (2006) Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15, 265-286.