Charles University

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Artificial Prediction Markets, Forecast Combinations and Classical Time Series

Author: **Bc. Marek Lipán**

Supervisor: **doc. PhDr. Jozef Baruník, Ph.D.**

Academic Year: **2017/2018**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, July 31, 2018

Signature

Acknowledgments

I would like to express my deepest gratitude to my supervisor
doc. PhDr. Jozef Baruník, Ph.D. for guidance, inspiration and willingness to discuss ideas and thoughts relating my thesis.

I would like to thank Mgr. Veronika Konečná and my family for the limitless moral support.

Abstract

Economic agents often face situations, where there are multiple competing forecasts available. Despite five decades of research on forecast combinations, most of the methods introduced so far fail to outperform the equal weights forecast combination in empirical applications. In this study, we gather a wide spectrum of forecast combination methods and reexamine these findings in two different classical economic times series forecasting applications. These include out-of-sample combining forecasts from the ECB Survey of Professional Forecasters and forecasts of the realized volatility of the U.S. Treasury futures log-returns. We asses the performance of artificial predictions markets, a class of machine learning methods, which has not yet been applied to the problem of combining economic times series forecasts. Furthermore, we propose a new simple method called Market for Kernels, which is designed specifically for combining time series forecasts. We found that equal weights can be significantly outperformed by several forecast combinations, including Bates-Granger methods and artificial prediction markets in the ECB Survey of Professional Forecasters application and by almost all examined forecast combinations in the financial application. We also found that the Market for Kernels forecast performance is comparable to the best forecast combinations from the literature in both of the applications.

JEL Classification	C00, C53, C58
Keywords	Forecast combinations, artificial prediction markets, Market for Kernels, forecasting economic time series

Author's e-mail	`98888094@fsv.cuni.cz`
Supervisor's e-mail	`barunik@fsv.cuni.cz`

Abstrakt

Ekonomičtí agenti se často dostávají do situací, kde mají k dispozici několik odlišných předpovědí. Navzdory pěti dekádám zkoumání kombinací předpovědí, většina metod, která byla zatím představena, nedokáže v empirických aplikacích významně porážet kombinaci předpovědí s rovnoměrnými váhami. V této studii dáváme dohromady široké spektrum kombinací předpovědí a přezkoumáváme tyto zjištění ve dvou různých aplikacích předpovídání klasických ekonomických časových řad. Tyto zahrnují mimo-výběrové kombinování předpovědí ECB Survey of Professional Forecasters a předpovědí realizované volatility logaritmických zisků futures na americké státní dluhopisy. Hodnotíme výkonnost umělých predikčních trhů, třídy metod ze strojového učení, která zatím nebyla aplikována na problém kombinování předpovědí ekonomických časových řad. Dále navrhujeme novou jednoduchou metodu nazvanou Market for Kernels, která je navržená speciálně pro kombinování předpovědí časových řad. Zjistili jsme, že rovnoměrné váhy se dají významně porazit několika kombinacemi předpovědí, které zahrnují Bates-Grangerovi metody a umělé predikční trhy v ECB Survey of Professional Forecasters aplikaci a skoro všemi zkoumanými kombinacemi předpovědí ve finanční aplikaci. Také jsme zjistili, že předpovědní výkonnost Market for Kernels v obou aplikacích je srovnatelná s nejlepšími kombinacemi v literatuře.

Klasifikace JEL	C00, C53, C58
Klíčová slova	Kombinace predikcí, umělé predikční trhy, Market for Kernels, předpovídání ekonomických časových řad

E-mail autora	`98888094@fsv.cuni.cz`
E-mail vedoucího	`barunik@fsv.cuni.cz`

Contents

List of Tables

List of Figures

Master's Thesis Proposal

Author	Bc. Marek Lipán
Supervisor	doc. PhDr. Jozef Baruník, Ph.D.
Proposed topic	Artificial Prediction Markets, Forecast Combinations and Classical Time Series

Motivation When rational economic agents look for optimal economic decisions under uncertainty, they often need to rely on outcomes of predictive modeling. In many cases, there is a set of considered prediction models from which a single model can be selected based on various criteria. However, assuming the considered models are reasonably diverse and their predictions accurate, a suitable combination of predictions of these models, i.e. so-called model ensemble, achieves more accurate results than any of the individual predictors (Dietterich, 2000).

Artificial prediction market is a form of a model ensemble, firstly presented by Lay & Barbu (2010). Its idea stems from the real prediction market, which is a mechanism for aggregating disperse information from a number of prediction market participants. These participants trade contracts with payoffs dependent on unknown future events (Wolfers & Zitzewitz, 2004). Storkey (2011) showed that various forms of aggregation can be achieved by using different utility functions of agents participating in artificial prediction market. Artificial prediction market learning algorithm has been applied in both classification problems e.g. (Barbu & Lay, 2012), (Millin et al., 2012) or (Hu & Storkey, 2014) and regression problems e.g. (Lay & Barbu, 2012), (Jahedpari et al., 2017). Economic agents or policy makers are typically interested in prediction of future evolution of certain macro or microeconomic variables. Researchers have been looking for optimal way to combine forecasts since the seminal work of Bates & Granger (1969). With the aim to contribute to the pool of literature on ensemble forecasting, I will extend the artificial prediction market methodology for market agents with new features, suited specifically for time series data and try to outperform in forecasting the benchmark model ensembles as well as single models, which are traditionally used for economic time series analysis.

Hypotheses

Hypothesis #1: The artificial prediction market forecasts better than the benchmark model ensembles.

Hypothesis #2: The artificial prediction market forecasts better than any of its single model components.

Hypothesis #3: The artificial prediction market with agents having time series specific features forecasts better than the basic artificial prediction market.

Methodology In my thesis, I will stick to artificial prediction market in the utility theory based framework presented by Storkey (2011). I will extend it from a classification problem to a regression problem in a similar fashion as Lay & Barbu (2012) or Jahedpari et al. (2017) did it for artificial prediction market in a betting functions framework. As individual model components I will use various econometric models for time series analysis and forecasting, including models from class ARIMA, exponential smoothing, regression, vector autoregression models and others. The list of benchmark model ensembles will include Bagging, AdaBoost and Random Forest, which are the ensemble methods heavily used in general machine learning problems, but also combining forecasts methods used in economic environments such as the Bates and Granger optimal combination, LASSO-based methods or the Bayesian Model Averaging (BMA). Similarly to Millin et al. (2012), I will experiment with new types of utility functions of artificial market participants to find the best to suit the time series forecasting problem. Most importantly, I will introduce agents with different investment horizons. The key idea is that different agents/models might be suitable for forecasting different frequency components of the given time series. The option of giving only certain parts of the frequency spectra to different agents as their available information input will be also considered.

The key part of my thesis will be the evaluation and comparison of models based on the quality of their forecasts. The accuracy of forecast is generally agreed upon to be the most important criterion for selection of the best forecasting methods (Yokuma & Armstrong, 1995). For comparing forecasts on single time series, error measures such as the most commonly used Root Mean Square Error (RMSE) will be employed. When selecting the best forecasting method across a set of series however, requirements for reliability, construct validity, protection against outliers and relationship to decision making arise and lowest RMSE is no longer a suitable criterion (Armstrong & Collopy, 1992). Following the guidelines of Armstrong & Collopy (1992) for comparing between the forecasting methods, I will employ the Median Relative Absolute Error (MdRAE) and the Median Absolute Percentage Error (MdAPE) measures. For the final assessment on the performance of the considered forecast ensembles,

the error measures will be complemented by formal tests of equal forecast accuracy proposed by Diebold & Mariano (2002) and tests of forecast encompassing discussed by Harvey & Newbold (1998).

Expected Contribution The goal of my thesis is to adapt artificial prediction market in the utility based framework so it can potentially become a legitimate member of the group of state of art forecast ensembles used in economics by researchers, policy makers or other economic agents, whose action depend on forecasts. Further, I will propose agents for artificial prediction market with time series specific features. These agents will have different types of utility functions, different investment horizons and different information sets determined based on spectral analysis. It will be examined whether these ideas present a possible way to improve the artificial prediction market for time series. Finally, I will compare the performance of both the basic artificial prediction market and the extended artificial prediction market with the benchmark ensembles for time series forecasting using various error measurements and series of tests.

Outline

1. Introduction
2. Literature review: review of the literature covering model ensembles, combination of forecasts in economics, prediction markets and artificial prediction markets
3. Methodology: presentation of methodology and idea behind individual forecasting models, benchmark forecast ensembles and artificial prediction markets
4. Data: description of economic time series used for modeling and forecasting
5. Forecast comparison and testing: assessment of forecast quality of considered ensembles using error measures, tests of equal forecast accuracy and tests of forecast encompassing
6. Discussion
7. Conclusions

Core bibliography

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. International journal of forecasting, 8(1), 69-80.

Barbu, A., & Lay, N. (2012). An introduction to artificial prediction markets for classification. Journal of Machine Learning Research, 13(Jul), 2177-2204.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. Or, 451-468.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. Journal of Business & economic statistics, 20(1), 134-144.

Dietterich, T. G. (2000). Ensemble methods in machine learning. Multiple classifier systems, 1857, 1-15.

Harvey, D. S., Leybourne, S. J., & Newbold, P. (1998). Tests for forecast encompassing. Journal of Business & Economic Statistics, 16(2), 254-259.

Hu, J., & Storkey, A. (2014). Multi-period trading prediction markets with connections to machine learning. In International Conference on Machine Learning (pp. 1773-1781).

Jahedpari, F., Rahwan, T., Hashemi, S., Michalak, T. P., De Vos, M., Padget, J., & Woon, W. L. (2017). Online Prediction via Continuous Artificial Prediction Markets. IEEE Intelligent Systems, 32(1), 61-68.

Lay, N., & Barbu, A. (2010). Supervised aggregation of classifiers using artificial prediction markets. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) (pp. 591-598).

Lay, N., & Barbu, A. (2012). The Artificial Regression Market. arXiv preprint arXiv:1204.4154.

Millin, J., Geras, K., & Storkey, A. J. (2012). Isoelastic agents and wealth updates in machine learning markets. In Proceedings of the 29th International Conference on Machine Learning (ICML-12) (pp. 1815-1822).

Storkey, A. J. (2011). Machine learning markets. In International Conference on Artificial Intelligence and Statistics (pp. 716-724).

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. The Journal of Economic Perspectives, 18(2), 107-126.

Yokuma, J. T., & Armstrong, J. S. (1995). Beyond accuracy: Comparison of criteria used to select forecasting methods. International Journal of Forecasting, 11(4), 591-597.

Chapter 1

Introduction

Many economic agents, including policy makers, bankers, risk analysts, investors and even households base their decisions on forecasts or projections of certain economic variables. Numerous researchers are thus encouraged to develop models and methods yielding accurate forecasts of economic time series. As a result, the agents striving for as accurate forecasts as possible often find themselves in situations where there are multiple competing hypotheses or models available, offering different forecasts. In such situations, there are in principle two options. The first is to assume that the true model of the underlying data generating process (DGP) of interest corresponds to one our considered model specifications. Then, the econometric literature offers variety of tests and information criteria, based on which it can be determined, which of the model specifications is most likely the true one. If it is agreed that one of the competing models is indeed the true model, then forecasts based on this model are optimal and the other forecasts can be discarded. The second option is to acknowledge that the true model specification is likely missing among available ones or that we do not have a procedure to reliably detect it at our disposal. This acknowledgement leaves the door open for the possibility that a suitable combination of forecasts could yield forecasts superior in accuracy to all the individual ones. As Bates & Granger (1969) note, if the individual forecasts contain independent information stemming from either different information sets or model specifications, there is a room for improving upon them by their combination. In a search for optimal forecasts, one should combine the information sets rather than forecasts (Engle *et al.*, 1984). However, consider for example the case of European Central Bank (ECB) Survey of Professional Forecasters (SPF), where individual forecasters from different agencies contribute by their point forecasts of macroeconomic variables. In such cases, there is no choice but to apply some sort of forecast combination technique. Since the seminal work of Bates & Granger (1969), the literature on forecast combinations has grown substantially in size. Nevertheless, we find that these forecast combinations

started to revolve around the same ideas and that there is for example a rather limited amount of empirical literature applying classical methods from the field of machine learning to the task of economic forecast combinations. Here we should explain, that by the forecast combination we understand essentially any method, which can be used to assign weights to individual forecasts and combine them into a single forecast. The machine learning methods are suitable for modelling arbitrary functional forms and due to their penalization and dimension reduction techniques they can deal even with large number of predictors (Gu *et al.*, 2018). This makes them natural candidates for ideal forecast combinations. The researchers often avoid using such methods in economic time series forecasting applications, because of the phenomenon called the forecast combination puzzle, which states that the simple forecast combinations such as the equal weights (simple average) tend to outperform the more sophisticated ones (Jeremy & F., 2009). Nevertheless, we see an unexplored opportunity in the artificial prediction markets, a recent stream of literature started by Lay & Barbu (2010) and inspired by the real prediction markets (Wolfers & Zitzewitz, 2004). These methods are quite simple, yet flexible, based on interesting ideas and could represent a good alternative to the traditional forecast combinations. We see appropriate to asses the performance of the artificial prediction markets forecast combinations against the more traditional ones.

In this study, we examine and compare the pseudo-out-of-sample (further referred to simply as the out-of-sample) performance of a wide spectrum of forecast combination methods in two different classical economic time series forecasting applications. The spectrum covers most of the forecast combinations presented in the literature so far, examples of the artificial prediction markets methods and a method called Market for Kernels, which is a new simple method we have designed for time series forecast combining. Our goal is to inspect whether any conclusions can be made about the performance of the examined forecast combinations in different forecasting environments, that could be of use to both academicians and practitioners, who wish to apply the forecast combination methods to economic time series. Additionally, we aim to examine whether the forecast combinations actually improve upon the individual forecasts in our empirical applications. The first application presented in this study is the combining of forecasts of macroeconomic variables at different horizons from the ECB SPF. In the second one, the financial application, we consider combining forecasts of the realized volatility of the U.S. Treasury futures log-returns. For the purpose of examination of the forecast performance, we use mainly the most common accuracy measures and a test of equal forecast accuracy by Diebold & Mariano (2002). In order to challenge the forecast combination puzzle phenomenon, we test the null hypothesis of equal forecast accuracy of the equal weights forecast combination and the other studied forecast combinations in both

applications. Furthermore we focus our attention on assessing the performance of the newly proposed method. We test the null hypothesis of equal forecast accuracy of the Market for Kernels and the other forecast combination methods in both applications. Lastly, we test the null hypothesis of equal accuracy of Market for Kernels and the individual volatility forecasting models considered in the financial application. The python scripts with our implementation of all the forecast combinations used as well as scripts we used to produce all the output tables and figures are publicly available at `https://github.com/MarekLipan/master_thesis_sc`.

The thesis is structured as follows. After the introduction, in the second chapter, we review the literature on forecast combinations including the artificial prediction markets methods. Then, in the third chapter, we in detail describe the methodology of the studied forecast combinations and present the Market for Kernels method. In the fourth chapter, we present both of our empirical applications, describe the data used and the individual realized volatility forecasting models from the financial application. Next, in the fifth chapter, we describe the accuracy measures, the test used and present the results from both of our empirical applications. Then, in the sixth chapter, we discuss our findings in context of the forecast combination literature and share some of our thoughts and suggestions regarding the topic. Finally, the chapter seven concludes.

Chapter 2

Combining Forecasts in the Literature

There is a plethora of methods suggested for combining forecasts in the literature. And there are multiple possible ways of how these methods could be sorted into different groups under various labels and in what order they could be presented, because there is no consensual approach in that matter in the literature. Some general reviews or sortings of forecast combination methods were done in e.g. De Menezes *et al.* (2000), Stock & Watson (2004), Timmermann (2006) or Genre *et al.* (2013), but neither of them encompasses all the methods mentioned here. Therefore a following simple way of sorting methods into sections was chosen. Firstly are presented the methods, which are usually the most simple ones in terms of implementation, understanding, computational intensity and most of them were hierarchically proposed among the first ones or were directly derived from them. Then, slightly more sophisticated methods are summarized including factor analytic (principal component) combinations, shrinkage methods and Bayesian model averaging techniques. Then is presented a selection of methods from the classes of bagging, boosting and artificial neural network methods. These methods are in this study referred to as the alternative methods, because they are vastly more applied in the field of general machine learning rather than the traditional economic forecast combination literature. Lastly follows the summary of all of the relevant literature from the field of artificial prediction markets presented so far. The artificial prediction markets methods are excluded from the class of the alternative (machine learning) methods and receive a special attention, because our proposed method, the Market for Kernels, is heavily inspired by this particular stream of literature and would itself emerge within the class.

2.1 Simple Forecast Combinations

The origins of combining economic forecasts are often attributed to Bates & Granger (1969). The authors consider a general case where there are multiple different forecasts of a given variable available. They argue that if the primary objective is not to analyze the relationships in the data but rather to achieve as good forecast as possible, it is advisable to consider a suitable combination of forecasts rather than to select and rely on only one of the forecasts. Bates & Granger (1969) recognize two reasons why a combined forecast might be more precise than each of the individual forecasts. Firstly, the individual forecasts can be based on different information. Secondly, there can be different relationships among variables assumed by the individual forecasts. The usefulness of combining forecasts is demonstrated on the problem of one period ahead forecasting of monthly passenger miles flown in 1953. It is shown that a simple average of the Brown's exponential smoothing forecast and the Box-Jenkins adaptive forecast yields a combined forecast with a lower mean square error than both of the individual forecasts. Bates & Granger (1969) then suggest five different methods in which two series of forecasts are linearly combined with the aim to minimize the mean square error of the composite forecast. The weights of individual forecasts are calculated based on the past performance of the respective forecasts. The methods make no assumptions about the underlying model. The authors explain that in case we assumed a single model specification is correct, the individual forecast based on this particular model could not be improved by combining it with forecasts based on other (incorrect) model specifications. An important conclusion drawn by Bates & Granger (1969) is that methods with changing weights usually lead to better forecasts than the methods with constant weights, which are calculated based on all of the observed errors from the sample.

In an empirical study by Newbold & Granger (1974) on a sample of 80 monthly economic time series, the assessment of forecasting univariate time series performance using 3 different methods (Box-Jenkins, Holt-Winters and stepwise autoregression) is made. From the considered methods, the Box-Jenkins method was found to perform the best. However, it requires more time and skill than Holt-Winters and stepwise autoregression methods, which are fully automatic. Importantly, Newbold & Granger (1974) found that by combining the two automatic methods using suitable weights computed as suggested by Bates & Granger (1969), a performance which is superior to both individual methods and closely approximates the one of Box-Jenkins method is achieved. Furthermore, the authors report that Box-Jenkins forecasting performance can be often improved upon when the forecasts of all the 3 methods are combined.

Additional empirical evidence suggesting that combining forecasts improved per-

formance is provided by Stock & Watson (1998a). In their study, 49 linear and non-linear univariate forecasting methods and various forecast combinations are examined on a dataset of 215 U.S. macroeconomic monthly time series. The findings are that even the best performing individual method (autoregressions with unit root pretest) can be improved upon by combining it with the other methods. Especially, as the most reliable combining methods have proven to be the ones that place weights on all of the individual methods (e.g. equal weighting or inverse MSE weighting).

Despite many economic studies find that combining forecast leads to improved forecast accuracy, Yang (2004) argues that combining forecasts blindly can drastically worsen the performance due to the large variability in estimating the combining weights. The author recognizes a potential gain from sharing the strengths of different individual forecasting models and at the same time the price of combining in terms of complexity. To achieve a good balance between these two, Yang (2004) proposes the Aggregated Forecast Trough Exponential Re-weighting (AFTER). This automated algorithm works with recursively updated weights based on the past performance of the individual forecasts. It has a property of achieving similar performance as the best one of the individual forecasts and is easy to implement. This property is very useful, because we often do not know apriori, which of the forecasts would perform the best. Yang (2004) concludes that combining the candidate forecasting models with weights assigned by AFTER leads to more stable predictions and a better performance than in the case the attempt is made to select the best of the candidate models based on information criteria.

Zou & Yang (2004) deal with an empirical problem of model selection when forecasting time series using ARIMA models. They find in a simulation study as well as on a real data example that selecting a single most appropriate model based on information criteria often leads to unstable results. Zou & Yang (2004) suggest instead to convexely combine some of the considered reasonable ARIMA models using the AFTER algorithm as proposed by Yang (2004), which shows to lead to improved performance.

Hendry & Clements (2004) show theoretically, how pooling forecasts leads to improvement in a case when no single model coincides with the DGP. They explain that averaging misspecified forecasts provides “insurance" and might provide dominance in case there is a location shift in some omitted variable. They even argue that averaging might lead to better performance than using estimated combining weights in such cases. For practical purposes, Hendry & Clements (2004) advise to robustify the combination by using median or trimmed mean to identify and discard outlying forecasts, which could otherwise have a bad impact on the whole combination.

Granger & Jeon (2004) discuss both theoretical and empirical benefits of thick modelling (i.e. keeping and synthesizing model outputs of close model specifications)

versus thin modelling (i.e. selecting and keeping output of only one of available model specifications). In a task such as forecast combining, it is often worth to not disregard the information available from alternative model specifications. Granger & Jeon (2004) suggest trimming $k\%$ of the lowest and highest forecasts and taking a simple average of the remaining ones. This procedure is very easy to implement as it requires no estimation of combining weights. By trimming, we only get rid of the portion of models which are most likely to be severely misspecified. Although, as the authors note, it might be sometimes the case in practice that forecasts coagulate around a value that is not really a good forecast. Using equal weights (simple average) has a lot of empirical support in the literature. The procedure suggested by Granger & Jeon (2004) thus appears simple, while effective at the same time.

Quite original approach to combining forecasts is suggested by Engle *et al.* (1984). There are two competing model forecasts of U.S. inflation considered in the study. Both are based on a different information set. One is a stylized monetarist model (St. Louis equation) and the other is based on markup pricing model. The authors propose to estimate a bivariate autoregressive conditional heteroskedasticity (ARCH) model of the forecast errors to determine the combining weights. The weights are time-varying and calculated from the conditional covariance matrix of forecast errors. If the forecasting performance of one of the considered models improves during some period of time (i.e. variance of the forecast errors reduces), the weight put on this particular model in the combination increases. Engle *et al.* (1984) note that observing the patterns in the time-varying weights can bring useful insights on how the descriptive ability of the models evolves over time or in what situations each model dominates the other.

Original and surprisingly simple approach is proposed by Capistrán & Timmermann (2009). The study is focused on a problem of forecasting data from various surveys of experts, which are often subject to entering and exiting of experts, leaving the researchers with unbalanced panels of data. Capistrán & Timmermann (2009) point out that this incompleteness in the data has large detrimental effect on the commonly used forecast combination methods and that it is ignored in the literature more than it should be. The proposed method is done in two stages. In the first stage, the equal weight forecast (simple average) from all available forecasts is obtained. In the second stage, the observed target variable of interest is regressed on a constant and the composite forecast obtained in the first step. It is shown that this kind of projection on equal weighted forecast leads to adjustment for biases and noise in the underlying aggregate forecast, which arise from the continual entering and exiting of experts from the survey. Using equal weights in this method is inspired by a general agreement on a good performance of a simple average in forecast combination tasks. The biggest advantage of the approach proposed here is that it

enables to utilize all the available data in comparison to combination methods that require estimation of a complete covariance matrix, which is not possible from unbalanced panels. The improved performance of the projection on equally weighted forecast method in comparison to other combination methods is demonstrated in a Monte Carlo simulation study and empirical application of forecasting inflation from the Survey of Professional Forecasters.

2.2 Factor Analytic Methods

Chan *et al.* (1999) consider forecast combinations in a dynamic factor model framework. They consider a panel of individual forecasts with a conditional expectation as a single unobserved factor. The approach here is to take the first principal component of the estimated second moment matrix of individual forecasts and then use it as the factor estimate. Then the parameter of this factor can be easily estimated using ordinary least square (OLS) regression. Their analysis is conducted on results of a Monte Carlo simulation experiments as well as empirical results from univariate forecasting of U.S. macroeconomic time series. The performance of the presented factor model is compared, based on the mean squared out-of-sample forecast error, with other combination methods. Although the main benefits of combining forecasts are often attributed to gains from pooling forecasts using different information sets, the authors report an interesting finding that combined forecasts led to solid improvement upon the individual forecasts, despite that all the considered individual forecasts in their example are univariate and are thus based on basically the same information set. While the principal component (factor analytic) method is shown to perform well in the simulation experiments, it is outperformed by other simpler methods such as the simple average, median or the trimmed mean in the empirical exercise.

In the paper by Stock & Watson (2004), the goal is to forecast quarterly output growth data from 7 OECD countries in the period 1959–1999. The authors work with 73 different predictors per each country. In situations such as this one, where there are very large number of predictors available in relative to the number of observations, approaches such as the simple OLS combining weights (suggested by Granger & Ramanathan (1984)) are inappropriate. The authors consider for this task several forecast combining methods, including principal component forecast combination. The mean squared forecast errors (MSFEs) of these methods are then compared to a dynamic factor model alternative. The suggested approach is similar to Chan *et al.* (1999). Firstly we compute the recursive individual forecasts and then obtain first m principal components of their matrix of uncentered second moments and so obtain the estimates of the common factors. Secondly, we regress the sample

of observed output growth on these factors using OLS and obtain the combining weights. Finally, we use the estimated weight to compute the combined forecast. Stock & Watson (2004) use 2 versions of this model with different number of principal components used. For the selection of number of principal components to be used, the authors suggest using either the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). There are several observation made from their empirical analysis. The dynamic factor model was generally outperformed by the forecast combination methods. More sophisticated forecast combination methods, especially those allowing for time varying weights, often perform poorly in comparison to the simpler methods, which also give more stable results.

Another example of using principal component (factor analytic) combining forecast method can be seen in a work of Genre *et al.* (2013). Here the authors use several forecast combination methods including the principal component method as suggested by Stock & Watson (2004). They use it to investigate whether there is a potential for improvement in combining quarterly macroeconomic variables forecasts from ECB Survey of Professional Forecasters. Genre *et al.* (2013) find that the relative performance of the examined method depends on both the forecasting horizon and the target variable. They found that there was not much of an improvement over the equal weights benchmark in forecasting GDP growth and unemployment, while there was some decent improvement in case of forecasting inflation. From the methods they examined, there was none which would generally dominate all the others across variables and horizons.

2.3 Shrinkage Methods

The first time the (Bayesian) shrinkage appears in the literature on forecast combinations is in the paper by Diebold & Pauly (1990). Here the authors propose a weighted average of OLS regression based combining weights and equal weights (simple average). As it is often noted, the equal weights, although theoretically suboptimal, provide a very solid benchmark, which is not easily beaten by other more sophisticated combining methods. Here the authors allow the weights to be shrunk towards but not force to be equal to simple average weights. The amount of shrinkage is driven by the strength of the prior on the weights, which can be estimated from the data using the empirical Bayes method proposed by the authors. Based on the empirical study of combining forecasts of U.S. GNP, Diebold & Pauly (1990) make several important observations. Firstly, all the individual predictors used in their example can be outperformed in forecasting precision by using the proposed combination method. Secondly, it is often necessary to shrink the weights by a relatively large amount. Thirdly, a simple average is the shrinking direction, which

provides the best final forecast combination. These findings are again indicative of a very good performance of a simple combining method such as the simple average.

Empirical applications of shrinkage methods can be also found in forecasting output growth of OECD countries by Stock & Watson (2004) and in forecast of macroeconomic variables from ECB Suvervey of Professional Forecasters by Genre *et al.* (2013). Stock & Watson (2004) in their application withhold from empirical Bayes estimation approach of their shrinkage parameter κ, in comparison to the approach of Diebold & Pauly (1990), because they deal with a very large number of predictors in their application. Instead they use several shrinking weights specified explicitly. As in Diebold & Pauly (1990), the shrinkage is done towards equal weights. Genre *et al.* (2013) directly follow the implementation of Stock & Watson (2004).

One of the most recent approaches towards the forecast combining is presented in Diebold & Shin (2017). The idea behind the 2-step egalitarian LASSO is inspired by the fact that a simple average usually performs very well. The reason is that equal weights always bring error variance reduction when combining forecast with uncorrelated errors. Also, while the equal weights are theoretically sub-optimal, they are likely to be close from optimal. Optimally, we want to give bigger weights to forecasts with lower variance of errors. Also, the 2-step egalitarian LASSO is designed to deal with a common situation where there are more individual forecast sets than actual observations. In the first step of the suggested procedure the usual LASSO, which "selects to 0", is run to discard some of the largely redundant forecasts and to determine k surviving ones. In the second step, the egalitarian ridge is used to shrink the remaining weights towards equality $(1/k)$. Diebold & Shin (2017) show that the 2-step egalitarian LASSO beats the simple average in terms of out-of-sample forecast RMSE (root mean square error) in forecasting Euro-area real GDP growth rate data from the European Central Bank's Survey of Professional Forecasters. As it is difficult to select the LASSO tuning parameter λ, which drives the strength of regularization, in the real-time with small samples, Diebold & Shin (2017) further propose the "best average" combinations method. This method is based on the lessons learned from egalitarian LASSO procedures and is free of any tuning parameters.

2.4 Bayesian Model Averaging Combinations

Jacobson & Karlsson (2004) explore the idea of using Bayesian Model Averaging (BMA) for combining forecasts of a large number of models. Their goal is to make a composite forecast of Swedish consumer price inflation index from a plenty of individual forecasts from models consisting of various possible combinations of about 80 available indicators. They suggest to use a BMA technique to determine the

posterior probabilities computed from marginal likelihoods of the models and then use it as combining weights. The results show that such model combination is quite robust and yields lower out-of-sample root mean squared forecast error than the individual models do. The findings of Jacobson & Karlsson (2004) are very supportive for the use of their methodology. Despite the persistence in the inflation rate, they successfully show that the BMA forecast combination can with ease outperform a random walk forecast in the 4 quarters horizon forecasting.

Eklund & Karlsson (2007) consider a Bayesian Model Averaging as an ideal framework for combining forecasts from the theoretical point of view. The BMA combinations have some nice properties grounded in the statistical theory, they can account for uncertainty in both models and parameters and can deal with situations where there are plenty of predictors available. Newly, Eklund & Karlsson (2007) propose to use the out-of-sample predictive likelihood to compute the combining weights and show why it is a better choice in that matter than the standard marginal likelihood. While the motivation is the same, the better fitting model from the candidate models should accumulate greater weight in the resulting combination, the weights computed based on the marginal likelihood are prone to in-sample overfitting. Eklund & Karlsson (2007) show that their new method still has good asymptotic properties. It converges to the true model, if it is present among the candidate ones. Although it converges at somewhat slower pace than when using the method with weights based on the marginal likelihood. Nevertheless, in practice, it can be hardly expected that the true model is among the candidate ones and in such situations the BMA forecast combination based on the predictive likelihood should provide the largest gains. Additionally, it has better small sample properties. Because it considers both in-sample fit and out-of-sample prediction, it provides some protection against overfitting in comparison to the standard marginal likelihood based weights. The arguments of authors are supported by their simulation study and the empirical exercise of forecasting Swedish inflation rate.

2.5 Alternative Methods

Donaldson & Kamstra (1996) propose to use artificial neural networks (ANN) as an instrument to combine forecasts. They explain that ANN should perform better than traditional linear forecast combination methods in situations where the optimal forecast combination is nonlinear. Based on the comparison of out-of-sample mean squared errors and tests of encompassing, Donaldson & Kamstra (1996) show that ANN outperforms in some of the traditional combining methods in one-step-ahead forecasting daily volatility of selected stock indices including S&P500, NIKKEI, TSEC and FTSE. As the individual forecasts for combining the authors used the

output from the MA variance model (MAV) and the GARCH(1,1) model. Harrald & Kamstra (1997) enrich the research on combining forecasts with ANNs and optimize the weights in the nodes of the ANN using the means of evolutionary programming. Donaldson & Kamstra (1999) follow up on their previous work and again demonstrate on the application of forecasting of S&P500 stock return volatility that ANN combination is superior to the traditional linear combination. Donaldson & Kamstra (1999) newly explain that the benefits of ANN combining lies in capturing the interaction effects between the individual forecasts. Their approach thus represents a way how to incorporate a state-dependency into forecast combination modelling.

We can also consider bootstrap aggregation (bagging) as another option of how forecasts can be combined. Bagging is a statistical technique designed to reduce out-of-sample mean squared forecast error in cases where there is a large number of predictors and where model selecting rules produce unstable results. Inoue & Kilian (2008) employ bagging with a simple pre-test strategy in a task of forecasting U.S. CPI inflation and compare the results with standard forecast combination methods, factor models as well as the bayesian model averaging method. Their results from the out-of-sample forecasting exercise indicate that bagging as used by Inoue & Kilian (2008) indeed is a good alternative to other forecast combining methods in case the data is covariance-stationary.

In a similar way to bagging, one can also use boosting to combine forecasts. Buchen & Wohlrabe (2011) use componentwise boosting to forecast U.S. industrial production growth at different horizons using 130 economic time series. The authors compare the performance of boosting, traditional forecast combinations and dynamic factor model and find that boosting works as a viable alternative. Adaptive Boosting (AdaBoost) is a boosting algorithm which was later generalized to Gradient Boosting (Friedman, 2001). Barrow & Crone (2016) asses the performance of AdaBoost against bagging and other combining methods in forecasting 111 monthly industrial time series from NN3 competition[1]. They examine multiple different versions of AdaBoost distinguished by the different choice of meta-parameters such as the loss function, the stopping criteria or the base model. All forecast combining procedures are found to be superior to individual best model selection. The authors introduce a novel algorithm called AdaBoost.BC, which employs the selection of the best AdaBoost meta-parameters for time series forecasting. Nevertheless, all the AdaBoost versions are shown to forecast less accurately than bagging or the other examined simpler forecast combination methods.

[1]More information about the NN3 competition is available on the following website http://www.neural-forecasting-competition.com/NN3/

2.6 Artificial Prediction Markets

The Artificial Prediction Markets is the method conceptually falling within the realm of the (real) Prediction Markets. According to Wolfers & Zitzewitz (2004), the prediction markets are financial markets where participants trade contracts whose payoffs depend on uncertain future outcomes. The main motivation relies on the efficient market hypothesis – in a an efficient market, the prices all the time fully reflect all the available information (Fama, 1970). In the fully efficient prediction markets, the market price represent the best possible predictor of the future events of interest. Wolfers & Zitzewitz (2004) distinguish three types of prediction markets: winner-take-all, index and spread. For this thesis are relevant the first two types. In the winner-take-all market, the contract costs e.g. $\$p$ and pays \$0 or \$1 depending on whether some future event of interest occurs or not. The price for contract p thus represents the market's expected probability of occurrence of the future event. In the index market, the payoff varies in accordance to some quantity such as e.g. percentage of votes obtained by a certain candidate in political elections. The market price thus represents mean value assigned by the market to the given quantity of interest. Wolfers & Zitzewitz (2004) recognize two base market designs. Firstly, a continuous double auction, where buyers put their bids and sellers asks and the market mechanism pairs the two sides whenever possible trade appears. Secondly, a pari-mutuel system, where market participants place bets in the common pot, which is later divided among the winning participants. The benefit of prediction markets is that they provide incentives for truthful revelation, information discovery and mechanism for aggregating opinions (Wolfers & Zitzewitz, 2004). An example of an existing application of the prediction markets is https://www.predictit.org/, a project of Victoria University of Wellington, where the participants trade the futures on political events. Apart from public prediction markets such as this one, many companies, including e.g. Microsoft, Siemens or Google, run their internal prediction markets in order to aggregate the information or beliefs about the future events dispersed among their employees Cowgill *et al.* (2009).

Artificial prediction markets were firstly introduced for aggregating classifiers by Lay & Barbu (2010). The market setup is inspired by a real prediction market – The Iowa Electronic Market, where the contracts for all possible outcomes of an event are sold and the one which correctly predicts the outcome pays \$1 after it is realized. The artificial prediction market participants are the individual classifiers. In the initialization step, the market participants are provided with equal budgets. In the training process, the participants allocate part of their budget for purchasing contracts according to their assigned betting functions. The betting functions can be of various types. Lay & Barbu (2010) introduce constant, linear and aggres-

sive betting functions. Also, they propose an algorithm for numerical derivation of the market equilibrium price of contracts. After the realized outcome is observed, the market participants are awarded in a budget updating procedure. This means that in a trained market, participants which are correct more frequently than the others accumulate greater share on a total budget and thus have a greater impact on the equilibrium price. The contract equilibrium price vector can be interpreted probabilistically as the aggregate classifier result.

Storkey (2011) introduces a utility-based framework to the artificial prediction markets or the machine learning markets as it is referred to in the paper. In contrary to the betting framework, here the market constitutes of a set of agents each with a defined utility function. Each agent acts as to maximize her utility function given a cost of traded goods, which are bets on individual outcomes. Depending on the type of the utility functions, market equilibrium price (fixed-point) can be derived, which can then be interpreted probabilistically. As Storkey (2011) notes, the machine learning markets represent a very flexible way of combining models. They show that various existing model combinations used in machine learning can be implemented using various utility functions and market designs. For example, a linear debt-utility gives weighted median model combination, a logarithmic utility gives a weighted mean model combination, exponential decaying negative utility gives a product model combination and so on. Moreover, the mechanism of the machine learning market as suggested by Storkey (2011) allows for integration of independent agents with different types of utility functions. This approach offers even greater versatility in model combining.

Millin *et al.* (2012) further build on the previous research of Storkey (2011). They introduce the agents with isoelastic utility functions and show how it can improve the market performance over the logarithmic and negative exponential utility functions. The market equilibrium cannot be computed analytically for a general elasticity parameter or in the case of inhomogeneous market (i.e. market consisting of agents with non-identical utility functions). Therefore the authors present an algorithm, which can be used to find the equilibrium numerically, based on the principle of minimizing the divergence of cost of the given good and the amount invested in it. Millin *et al.* (2012) consider 2 possible wealth updating schemes: online and batch, which correspond to bayesian model updates and mixing coefficient updates respectively. The wealth update mechanisms are vital for machine learning markets as they ensure that the weights are properly distributed among the agents in the training phase. Millin *et al.* (2012) demonstrate that inhomogeneous markets of isoelastic agents outperform some of the state of art classifiers on a number of UCI datasets[2], which are often used for machine learning benchmarking.

[2]The UCI datasets are available at https://archive.ics.uci.edu/ml/datasets.html

After the research of Storkey (2011) and Millin *et al.* (2012), Hu & Storkey (2014) choose different approach and consider agents whose decisions are driven by risk measures. In contrary to a general utility function, the risk measures satisfy a property of translation invariance, which implies that an optimal portfolio of a particular agent does not depend on her wealth. Hu & Storkey (2014) abandon interpreting the agent's wealth as the aggregating weights, because they find the relationships among them are very inconsistent and vary greatly based on which utility functions are used. Instead, they propose a framework in which agents trade with a market maker in a multi-period trading scheme. While each agent separately is following her own goals, the market is analytically shown to be optimizing a certain global objective. This allows for establishing a direct connection between the market and machine learning. Machine learning problems of some form can be transformed into a market and its solution find by running the market. As the examples Hu & Storkey (2014) include opinion pooling, bayesian updates and logistic regression.

The first adaptation of artificial prediction market to a regression problem is done by Lay & Barbu (2012). Because there are uncountably many possible outcomes in a regression problem, the reward kernel is introduced, which is a density centered around the true value of the dependent variable. The reward kernel determines the size of the reward for each agent after the true value is observed. For each given outcome from the space of possible values of the dependent variable, the reward for a bet at that given outcome is a function of the absolute difference of the given outcome and the true value (prediction error). The lower the difference, the greater the reward. The wealth updating rules depend on a selection of the reward kernel. Lay & Barbu (2012) present delta updates and Gaussian updates. The authors show that their artificial regression market significantly outperforms a random forest regression on a number of UCI datasets.

Most recently, Jahedpari *et al.* (2017) propose a continuous artificial prediction market (c-APM) for online regression problems, in which the market is trained on observations one by one. In c-APM, the pari-mutuel mechanism extended for regression is used, in which market participants bet on outcomes of their choice and the aggregated prediction is taken as the average of these outcomes weighted by sizes of the bets. The most important innovation in a work of Jahedpari *et al.* (2017) over the previous research is that the agents are allowed to have adaptive strategies and can revise their bets based on how bet the other agents in a pari-mutuel mechanism in multiple rounds and thus incorporate “the wisdom of the crowd". The authors examine 2 trading strategies: constant trading and Q-learning, which is a reinforcement learning technique used for finding an optimal action-selection policy. The c-APM is shown to predict very well when compared with other prediction aggregating models and the individual predictors, when tested on UCI datasets.

Chapter 3

Methodology of Forecast Combinations

As discussed in the beginning of the literature review 2, there is no consensual approach on how the individual forecast combining methods should be categorized into labelled classes. Nor we believe there is a necessity for making a unified approach or that it is even possible. The reason is that one can always create a method by combining features of methods from different classes or come up with entirely different approach to solving the forecast combining problem, which would not convincingly fit into any of the previously established classes. As it was already explained in 2, in dividing combination methods between sections, we are guided by the principles of common idea, complexity, relevance and time hierarchy. The chapter is concluded by the newly proposed method, the Market for Kernels, which is presented within the class of artificial prediction markets methods.

3.1 Simple Forecast Combinations

In this largest class of the forecast combination methods presented in this study is the traditional benchmark from the forecast combination literature – the equal weights forecast, the original Bates-Granger combinations as well as other traditional combining methods and methods, which we find relatively simple enough to understand, implement and compute.

3.1.1 Equal Weights

Many researchers report in their empirical works that simple combination methods such as the equally weighted (simple average) forecast represents a benchmark, which is tough to beat in the forecast accuracy by other, more sophisticated methods

(e.g. Stock & Watson (2004), Genre *et al.* (2013) or Conflitti *et al.* (2015)). This phenomenon is known in the literature as the forecast combination puzzle (Smith & Wallis, 2009). Smith & Wallis (2009) show that in case when the optimal weights are close to equality, the simple average is expected to give lower mean square forecast errors (MSFEs) than other methods using estimated weights, due to the estimation variance. Claeskens *et al.* (2016) further develop the underlying theory and illustrate how estimation of optimal weights leads to bias and increase of variance of the forecast combination. This helps to explain why the simplest combination method available, the equal weights combination, although sub-optimal, might perform very well in comparison with other combination methods in empirical applications. The equal weights forecast combination at time T is defined as follows:

$$f_{C,T} = \frac{1}{K} \sum_{i=1}^{K} f_{i,T}, \tag{3.1}$$

where K is the number of available individual forecasts $f_{1,t}, \ldots, f_{K,t}$.

3.1.2 Bates-Granger Optimal Combining Weights

Bates & Granger (1969) present the original idea to combine a pair of forecasts of a certain variable to obtain a single combined forecast with a lower variance of forecast error than both of the individual forecasts. It is necessary to assume that the individual forecasts are unbiased in order to obtain an unbiased combined forecast. Bates & Granger (1969) believe that an ideal combination method should have the following three properties. Firstly, the weights should approach the optimal values as the number of forecasts increases. Secondly, as the relative accuracy of the individual forecasts evolves, the weights should adapt correspondingly. Lastly, the weights should not deviate too much from the optimal values. Bates & Granger (1969) propose five combination methods having these properties, while being simple enough, so that they can be easily applied in practice. These methods were, however, described only for a pair of individual forecasts. Here, we present the generalization of these five methods for a set of K forecast, in a form they were presented later by Newbold & Granger (1974).

Let's assume the following linear combination of forecasts at time T:

$$f_{C,T} = \boldsymbol{\omega}_T' \boldsymbol{f}_T, \quad \boldsymbol{\omega}_T' \mathbf{1} = 1, \quad 0 \leq \omega_{i,T} \leq 1 \quad \text{for } \forall i, \tag{3.2}$$

where $\boldsymbol{f}_T = (f_{1,T}, \ldots, f_{K,T})'$ is the vector of the individual forecasts, $\boldsymbol{\omega}_T' = (\omega_{1,T}, \ldots, \omega_{K,T})$ is the vector of weights and $\mathbf{1} = (1, \ldots, 1)'$ is the vector of ones. Let's further assume a vector of forecast errors $\boldsymbol{e}$:

$$\boldsymbol{e}_T = y_T \mathbf{1} - \boldsymbol{f}_T, \tag{3.3}$$

where y_T is the realization of the variable to be forecast at time T. In what follows are defined the weights used in the individual methods.

Bates-Granger (1)

$$\omega_{i,T} = \frac{\left(\sum_{t=T-\nu}^{T-1} e_{i,t}^2\right)^{-1}}{\sum_{j=1}^{K}\left(\sum_{t=T-\nu}^{T-1} e_{j,t}^2\right)^{-1}}, \tag{3.4}$$

where ν is the parameter controlling the length of the window used to calculate weights. We set the ν equal to T, in order to utilize all the available training sample data.

Bates-Granger (2)

$$\boldsymbol{\omega}_T = \frac{\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}, \quad \text{s.t. } 0 \leq \omega_{i,T} \leq 1 \quad \text{for } \forall i \tag{3.5}$$

where

$$(\hat{\boldsymbol{\Sigma}})_{i,j} = \nu^{-1} \sum_{t=T-\nu}^{T-1} e_{i,t}e_{j,t}.$$

This method utilizes the theoretical knowledge that the weights

$$\boldsymbol{\omega}_T = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}$$

minimize the error variance of the combined forecast (Newbold & Granger, 1974). In most of the empirical applications, however, the covariance matrix $\boldsymbol{\Sigma}$ must be estimated. The first method (3.4) is the special case of this method, where there are all the correlations between the individual forecasts assumed to be zero. In case some of the calculated weights fall out the interval $[0, 1]$, it is possible to replace them by the appropriate end points (Granger & Newbold, 1986). Again, we set ν equal to T in order to use all the available data.

Bates-Granger (3)

$$\omega_{i,T} = \alpha\omega_{i,T-1} + (1-\alpha)\frac{\left(\sum_{t=T-\nu}^{T-1} e_{i,t}^2\right)^{-1}}{\sum_{j=1}^{K}\left(\sum_{t=T-\nu}^{T-1} e_{j,t}^2\right)^{-1}}, \quad 0 < \alpha < 1. \tag{3.6}$$

The third method convexly combines the weights from the first method (3.4) and the directly preceding weights. The parameter α influences the pace of the weight

adaptation. We use α equal to 0.6, and as in the previous two methods, the rolling window length ν equal to T.

Bates-Granger (4)

$$\omega_{i,T} = \frac{\left(\sum_{t=1}^{T-1} W^t e_{i,t}^2\right)^{-1}}{\sum_{j=1}^{K}\left(\sum_{t=1}^{T-1} W^t e_{j,t}^2\right)^{-1}}, \quad W \geq 1. \tag{3.7}$$

The fourth methods resembles the first method (3.4), except that now instead of controlling the length of the window, it is put exponentially more weight on the more recent forecast errors using the parameter W. In our applications, we use the W equal to 1.5.

Bates-Granger (5)

$$\boldsymbol{\omega}_T = \frac{\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}, \quad \text{s.t. } 0 \leq \omega_{i,T} \leq 1 \quad \text{for } \forall i \tag{3.8}$$

where

$$(\hat{\boldsymbol{\Sigma}})_{i,j} = \frac{\sum_{t=1}^{T-1} W^t e_{i,t} e_{j,t}}{\sum_{t=1}^{T-1} W^t}, \quad W \geq 1.$$

The last method resembles the the second method (3.5), except that the weighting idea is used instead of window controlling as in the fourth method (3.7). Here, we also apply the weighting parameter W equal to 1.5.

3.1.3 Granger-Ramanathan Combining Weights

Granger & Ramanathan (1984) extend the literature on linear forecast combination with 3 additional methods. The combining weights in all of these methods can be obtained using the ordinary least squares (OLS) estimator.

Granger-Ramanathan (1)

The first suggested procedure is to regress the variable to be forecast y_t on the individual forecasts $f_{1,t}, \ldots, f_{K,t}$ without a constant, i.e. estimating the model:

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.9}$$

where $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K)$ is the $(T \times K)$ matrix of K individual forecasts, T is the length of the data sample, $\boldsymbol{\beta}$ is the $(K \times 1)$ vector of combining weights and $\boldsymbol{\epsilon}$ is the

$(T \times 1)$ vector of forecast errors. The combined forecast is then calculated as:

$$\boldsymbol{f}_C = \boldsymbol{F}\hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated combining weights from regression (3.9). The sufficient conditions for $f_{C,t}$ to be an unbiased estimator of y_t are:

(i) each of the individual forecasts $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K$ is unbiased

(ii) the combining weights sum up to one (i.e. $\mathbf{1}'\hat{\boldsymbol{\beta}} = 1$)

However, it cannot be expected that these conditions will generally hold in practice (Granger & Ramanathan, 1984).

Granger-Ramanathan (2)

The second method proposed by Granger & Ramanathan (1984) differs from the first method in a way that the constraint is imposed on the combining weights $\hat{\boldsymbol{\beta}}$ such that their sum equals one (i.e. $\mathbf{1}'\hat{\boldsymbol{\beta}} = 1$). This constraint ensures that the unbiasedness of individual forecasts $f_{1,t}, \ldots, f_{K,t}$ implies unbiasedness of the combined forecast $f_{C,t}$. The weights can be found as a solution to the following minimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta})^{\intercal}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta}) + 2\lambda(\mathbf{1}'\boldsymbol{\beta} - 1). \tag{3.10}$$

Computationally equivalent solution can be found by regressing $(y_t - f_{K,t})$ on $(f_{1,t} - f_{K,t}), \ldots, (f_{K-1,t} - f_{K,t})$ without a constant:

$$\boldsymbol{y}^* = \boldsymbol{F}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{3.11}$$

where $\boldsymbol{y}^* = (\boldsymbol{y} - \boldsymbol{f}_K)$, $\boldsymbol{F}^* = (\boldsymbol{f}_1 - \boldsymbol{f}_K, \ldots, \boldsymbol{f}_{K-1} - \boldsymbol{f}_K)$ is the $(T \times K - 1)$ modified matrix of individual forecasts, $\boldsymbol{\beta}^*$ is the $(K - 1 \times 1)$ vector of first $K - 1$ weights. The estimate of the last combining weight $\hat{\beta}_K$ of the forecast $f_{K,t}$ can be found as:

$$\hat{\beta}_K = 1 - \mathbf{1}'\hat{\boldsymbol{\beta}}^*,$$

where $\hat{\boldsymbol{\beta}}^*$ is the estimate of the parameter vector $\boldsymbol{\beta}^*$ from the regression (3.11). As in the first method, the combined forecast is then computed as:

$$\boldsymbol{f}_C = \boldsymbol{F}\hat{\boldsymbol{\beta}}.$$

Granger-Ramanathan (3)

Finally, Granger & Ramanathan (1984) suggest an unrestricted linear combination with a constant term. The weights can be obtained by regressing y_t on $f_{1,t}, \ldots, f_{K-1,t}$

with a constant. This approach is claimed to be the best of all three proposed ones, because the resulting combined forecast is unbiased even when the individual forecasts are biased. Also, such combined forecast can be shown to have the lowest mean squared error from all the three methods (Granger & Ramanathan, 1984). The assumed regression model is:

$$\boldsymbol{y} = \alpha\mathbf{1} + \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.12}$$

where α is the intercept, $\mathbf{1}$ is the $(T \times 1)$ vector of ones and the rest of the notation is the same as in the first method (3.9). The combined forecast can be obtained as:

$$\boldsymbol{f}_C = \hat{\alpha}\mathbf{1} + \boldsymbol{F}\hat{\boldsymbol{\beta}},$$

where $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are the estimates of parameters from the regression (3.12).

3.1.4 AFTER

Aggregated Forecast Trough Exponential Re-weighting (AFTER) theoretically described by Yang (2004) is another method in the class of simple linear forecast combinations. Yang (2004) presents several variations of the AFTER algorithm. All but one of the methods require and incorporate either the known conditional variance of the variable to be forecast y, estimates of the variance by the individual forecasting procedures or the estimates of the conditional distribution. The last method, which is also presented here, does not require it and is therefore suitable for instances where y is non-stationary or where we are only provided with the individual point forecasts to be combined. The combining weight $\omega_{i,T}$ for the i-th forecast at time T is computed as follows:

$$\omega_{i,T} = \frac{\pi_i \exp\left(-\lambda \sum_{i=1}^{T-1} \psi(e_{i,t})\right)}{\sum_{j=1}^{K} \pi_i \exp\left(-\lambda \sum_{i=1}^{T-1} \psi(e_{i,t})\right)}, \tag{3.13}$$

where π_i is the prior weight, K is the number of available individual forecasts, λ is a small enough tuning parameter (set equal to 0.15 in our applications), $e_{i,t}$ is the forecast error and ψ is a non-negative convex loss function, which does not need to be symmetric around zero. Similarly to Zou & Yang (2004), we set the prior weights equal to $1/K$ and use the square loss function. The equation (3.13) thus reduces to:

$$\omega_{i,T} = \frac{\exp\left(-\lambda \sum_{i=1}^{T-1} (e_{i,t})^2\right)}{\sum_{j=1}^{K} \exp\left(-\lambda \sum_{i=1}^{T-1} (e_{i,t})^2\right)}. \tag{3.14}$$

The combined forecast at time T is then computed as:

$$f_{C,T} = \boldsymbol{\omega}' \boldsymbol{f}_T,$$

where $\boldsymbol{f}_T = (f_{1,T}, \ldots, f_{K,T})'$ is the vector of the individual forecasts and $\boldsymbol{\omega}' = (\omega_1, \ldots, \omega_K)$ is the vector of combining weights.

3.1.5 Median Forecast

Another simple but useful approach to combining a set of individual forecasts is to put all weight on the median forecast. Hendry & Clements (2004) suggest it as a way to neglect the bad influence of outlying forecasts on the combination. They argue that fixed weights combinations may dominate over estimated combinations in situations where there is a location shift in the underlying DGP. It is because previously successful individual forecasting procedures can become very inaccurate after the shift. The median forecast combination at time T is defined as follows:

$$f_{C,T} = \text{median}\{f_{i,T}\}_{i=1,\ldots,K}, \tag{3.15}$$

where K is the number of individual forecasts.

3.1.6 Trimmed Mean Forecast

Granger & Jeon (2004) suggest as a part of their "thick modelling" procedure to trim a certain percentage of the lowest and highest forecasts. This step taken in order to remove the potential harmful effect of outlying forecasts. Further, they suggest using the equal weights forecast combination 3.1.1 of the remaining forecasts. Advantages of this method are its robustness and no requirement for the estimation, which is useful in situations with many individual forecasts or small samples. The α-trimmed mean forecast can be obtained as follows:

$$f_{C,T} = \frac{1}{K - 2\lfloor \alpha K \rfloor} \sum_{i=1+\lfloor \alpha K \rfloor}^{K-\lfloor \alpha K \rfloor} f_{(i),T}, \tag{3.16}$$

where K is the number of individual forecasts, $\alpha \in [0, 0.5)$ is the trimming parameter and $f_{(i),T}$ denotes the i-th order statistic. In the empirical applications, we report the results for α set equal 0.05.

3.1.7 PEW

Projection on equal weights (PEW) method was firstly presented by Capistrán & Timmermann (2009). It is based on the knowledge of the good performance of the equal weights forecast 3.1.1 and is designed for unbalanced panel data. The idea is to project the variable to be forecast $\boldsymbol{y}$ on the the simple average of the of the individual forecasts using the least squares method with an intercept, which is added in order to remove any bias of the equal weights forecast. Parameters of the following model are estimated:

$$\boldsymbol{y} = \alpha \mathbf{1} + \beta \left(\frac{1}{K} \sum_{i=1}^{K} \boldsymbol{f}_i \right) + \boldsymbol{\epsilon}. \tag{3.17}$$

And then the combined forecast for the period T can be computed as follows:

$$f_{C,T} = \hat{\alpha} + \hat{\beta} \left(\frac{1}{K} \sum_{i=1}^{K} f_{i,T} \right),$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimates of the regression model parameters.

3.2 Factor Analytic Methods

Principal component (factor analytic) forecast, presented by Chan *et al.* (1999), are based on the assumption that the panel of forecasts follows the factor model. This method exploits the factor structure and is suitable for cases, when there is a high number (K) of highly correlated individual forecasts as it provides a way to reduce the dimensionality and can deal with the multicollinearity. Considering one-step-ahead forecasts, we have the following factor model representation:

$$y_{t+1} = \mu_t + \epsilon_{t+1}, \tag{3.18}$$

$$f_{i,t+1} = \lambda_i \mu_t + e_{i,t} \quad \text{for } \forall i, \tag{3.19}$$

where y_{t+1} is the variable to be forecast, μ_t is the unobserved factor, ϵ_{t+1} is the innovation, $f_{i,t+1}$ is the i-th individual forecast, $\lambda_i = 1 + Cov(\psi_{i,t} + \nu_{i,t}, \mu_t)/Var(\mu_t)$, where $\psi_{i,t}$ and $\nu_{i,t}$ are the i-th model specification and estimation errors respectively. Finally, $e_{i,t}$ is the error term, which is uncorrelated with μ_t (Chan *et al.*, 1999).

Equation (3.19) can be suitably rewritten in a matrix form as:

$$\boldsymbol{f}_{t+1} = \boldsymbol{\lambda}\mu_t + \boldsymbol{e}_t, \tag{3.20}$$

where $\boldsymbol{f}_{t+1} = (f_{1,t+1}, \ldots, f_{K,t+1})'$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)'$ and $\boldsymbol{e}_t = (e_{1,t}, \ldots, e_{K,t})'$.

3.2.1 Principal Components Forecast

The first step of the principal component forecast method by Chan *et al.* (1999) is to estimate the unobserved common factor μ_t by the first principal component:

$$\hat{\mu}_t = \hat{\Lambda}'\boldsymbol{f}_{t+1}, \tag{3.21}$$

where $\hat{\Lambda}$ is the eigenvector of the matrix of second moments of individual forecasts $T^{-1}\sum_{t=0}^{T-1}\boldsymbol{f}_{t+1}\boldsymbol{f}'_{t+1}$ with the largest eigenvalue. Stock & Watson (1998b) show that $\hat{\mu}_t$ is a consistent estimator of μ_t under general conditions. The second step is the estimation of the following regression model by OLS:

$$y_{t+1} = \beta\hat{\mu}_t + \xi_t, \tag{3.22}$$

which is shown to give forecasts that are asymptotically efficient (Chan *et al.*, 1999). According to Stock & Watson (1998b), there is no necessity for using more than the first principal component in the regression, when K is sufficiently large. The final principal component combination forecasts for period T is given by:

$$f_{C,T} = \hat{\beta}\hat{\Lambda}'\boldsymbol{f}_T,$$

where $\hat{\beta}$ is estimate of the coefficient from the regression (3.22).

3.2.2 Principal Components Forecast AIC/BIC

The principal component forecast combination method suggested by Stock & Watson (2004) is similar to that of Chan *et al.* (1999). The difference is that Stock & Watson (2004) consider not one but m common factors in the equation (3.20). Estimates of these factors are again obtained as principal components:

$$\hat{\mu}_{i,t} = \hat{\Lambda}'_i\boldsymbol{f}_{t+1} \quad \text{for } i = 1, \ldots, m, \tag{3.23}$$

where $\hat{\Lambda}_i$ is the eigenvector of the matrix of second moments of the panel of individual forecasts corresponding to the i-th largest eigenvalue. The factor weights are then

estimated using OLS:

$$y_{t+1} = \beta_1 \hat{\mu}_{1,t} + \ldots + \beta_m \hat{\mu}_{m,t} + \xi_t. \tag{3.24}$$

Stock & Watson (2004) suggest two different ways how to select the parameter m. First option is to use the Akaike information criterion (AIC), which is in the case of OLS with normally distributed errors defined as follows:

$$AIC = T\log(\hat{\sigma}^2) + 2P, \tag{3.25}$$

where T is the length of the sample, P is the number of model parameters and $\hat{\sigma}^2$ is the biased estimator of the error variance:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{i=1}^{T} \hat{\epsilon}_i^2, \tag{3.26}$$

where $\hat{\epsilon}_i$ is the i-th model residual (Burnham & Anderson, 2004). The goal is to select such m from the set $\{1, 2, 3, 4\}$ (as suggested by Stock & Watson (2004)) that the AIC from the model (3.24) is minimized. Second option is to use the Bayesian information criterion, which is in the case of OLS with normally distributed errors defined as:

$$BIC = T\log(\hat{\sigma}^2) + P\log(T), \tag{3.27}$$

where again we select such m from the set $\{1, 2, 3, 4\}$ that the BIC from the model (3.24) is minimized. Finally, the resulting principal component combination forecast for period T can be obtained as:

$$f_{C,T} = \hat{\beta}_1 \hat{\Lambda}_1' \boldsymbol{f}_T + \ldots + \hat{\beta}_m \hat{\Lambda}_m' \boldsymbol{f}_T = (\hat{\beta}_1 \hat{\Lambda}_1' + \ldots + \hat{\beta}_m \hat{\Lambda}_m') \boldsymbol{f}_T,$$

where $\hat{\beta}_1, \ldots, \hat{\beta}_m$ are estimates of the coefficients from the regression (3.24).

3.3 Shrinkage Methods

Shrinkage estimators are naturally well suited for the forecast combination problem. Many of the previously described methods are based either on the experience with a good empirical performance of the equal weights combination (section 3.1) or the theoretical properties of the Bates-Granger optimal combination (section 3.1.2). The shrinkage methods allow us to obtain weights somewhere in between these two, depending on the data. In some cases, shrinkage estimators can also be used to remove the redundant individual forecasts by shrinking their weights to zero.

3.3.1 Empirical Bayes Estimator

The first example of using a shrinkage in a context of forecast combination is presented in Diebold & Pauly (1990). The authors base their idea on the evidence of a good performance of the equal weights estimator in the empirical literature and notoriously known optimal combining weights by Bates & Granger (1969) extended by Granger & Ramanathan (1984). Here we present only one of the two the methods suggested by Diebold & Pauly (1990), which is the empirical Bayes estimator. The advantage of the empirical Bayes estimator over the g-prior estimator is that it does not require specifying of the parameter g, since the prior precision can be estimated directly from the data. In the empirical Bayes estimator, the combining weights are shrunken from the unconstrained OLS combining weights $\hat{\boldsymbol{\beta}}$ (3.12) towards the equal weights $\boldsymbol{\beta}_0$ (3.1).

Let's consider a linear forecast combination:

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$ is the normally distributed error term. Let's further assume the normal prior on $\boldsymbol{\beta}$:

$$P(\boldsymbol{\beta}|\sigma) = \mathcal{N}(\boldsymbol{\beta}_0, \tau^2\boldsymbol{A}^{-1}),$$

where $\boldsymbol{A}$ is the precision matrix and the parameter τ controls the variance of the prior. By combining the prior with the likelihood function:

$$L(\boldsymbol{\beta}, \sigma|\boldsymbol{y}, F) \propto \sigma^{-T}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta})\right)$$

we obtain the normal posterior:

$$P(\boldsymbol{\beta}|\sigma, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\beta}_1, (\tau^{-2}\boldsymbol{A} + \sigma^{-2}\boldsymbol{F}'\boldsymbol{F})^{-1}),$$

where the posterior mean can be expressed as:

$$\boldsymbol{\beta}_1 = (\tau^{-2}\boldsymbol{A} + \sigma^{-2}\boldsymbol{F}'\boldsymbol{F})^{-1}(\tau^{-2}\boldsymbol{A}\boldsymbol{\beta}_0 + \sigma^{-2}\boldsymbol{F}'\boldsymbol{F}\hat{\boldsymbol{\beta}}).$$

By substituting $\boldsymbol{A} = \boldsymbol{F}'\boldsymbol{F}$ and replacing σ^2 and τ^2 with the estimators:

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{F}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{F}\hat{\boldsymbol{\beta}})}{T},$$

$$\hat{\tau}^2 = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{\mathrm{tr}(\boldsymbol{F}'\boldsymbol{F})^{-1}} - \hat{\sigma}^2,$$

we can obtain the empirical Bayes combining weights (Diebold & Pauly, 1990):

$$\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_0 + \left(1 - \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\tau}^2}\right)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \tag{3.28}$$

Finally, the forecast combination is computed as:

$$\boldsymbol{f}_C = \boldsymbol{F}\hat{\boldsymbol{\beta}}_1.$$

3.3.2 Kappa-Shrinkage

Another example of a usage of a shrinkage method for forecast combination, here referred to as the Kappa-Shrinkage, can be seen in Stock & Watson (2004) and Genre *et al.* (2013). The shrinkage of weights is done from OLS combining weights without an intercept $\hat{\boldsymbol{\beta}}$ (3.9) towards the equal weights $\boldsymbol{\beta}_0$ (3.1). Considering one-step-ahead forecasts, the combining weights can be computed as:

$$\boldsymbol{\omega} = \lambda\hat{\boldsymbol{\beta}} + (1 - \lambda)\boldsymbol{\beta}_0, \tag{3.29}$$

where the shrinkage weight λ is defined as:

$$\lambda = \max\left\{0, 1 - \kappa\left(\frac{K}{\nu - 1 - K}\right)\right\}, \tag{3.30}$$

where K is the number of individual forecasts, ν is the length of the the training sample and κ is the parameter, which drives the amount of shrinkage. The larger the κ, the higher the shrinkage towards equal weights. In our empirical applications, we set κ equal to 0.5. Finally, The combined forecast at time T is computed as:

$$f_{C,T} = \boldsymbol{\omega}'\boldsymbol{f}_T,$$

where $\boldsymbol{f}_T = (f_{1,T}, \ldots, f_{K,T})'$ is the vector of the individual forecasts and $\boldsymbol{\omega}' = (\omega_1, \ldots, \omega_K)$ is the vector of combining weights. The advantage of this method is that it is suitable for instances where there is a large number of individual forecasts relative to the sample size (Stock & Watson, 2004). The disadvantage is that the resulting combination weights largely depend on the parameter κ, which needs to be pre-specified.

3.3.3 2-step Egalitarian LASSO

Diebold & Shin (2017) propose several of Egalitarian LASSO based procedures including the 2-step Egalitarian LASSO, which is described here. It was chosen over the other procedures, as we believe it spells out most the original ideas of Diebold

& Shin (2017). The key property of the 2-step Egalitarian LASSO is that it selects to zero and shrinks towards equality. Also, the regularization property makes it a usable method in situations where there is a greater amount of individual forecasts than the sample length. The set of combining weights is obtained in the following two steps.

In the first step, a standard LASSO (least absolute shrinkage and selection operator) is used to determine which of the K individual forecasts will be used in the second step and which will be discarded. The LASSO estimator is defined as:

$$\hat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{i,t} \right)^2 + \lambda_1 \sum_{i=1}^{K} |\beta_i| \right), \tag{3.31}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ is the $(K \times 1)$ vector of weights, T is the length of the sample and λ_1 is the tuning parameter, which governs the amount of shrinkage. Those k individual forecasts, for which the estimated weight is non-zero, are then used in the second step.

In the second step, the combining weights for the remaining k individual forecasts (denoted as $\boldsymbol{f}_1^*, \ldots, \boldsymbol{f}_k^*$) are estimated using Egalitarian LASSO, which in oppose to the standard LASSO shrinks towards simple averages (equality). It is defined as follows:

$$\hat{\boldsymbol{\beta}}_{EgalLASSO} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k} \beta_i f_{i,t}^* \right)^2 + \lambda_2 \sum_{i=1}^{k} \left| \beta_i - \frac{1}{k} \right| \right), \tag{3.32}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is the $(k \times 1)$ vector of weights and λ_2 is the tuning parameter, which is generally different from λ_1 in (3.31).

The combining weights produced by this procedure largely depend on the selection of the tuning parameters λ_1 and λ_2. Diebold & Shin (2017) suggest to use the ex ante optimal tuning via leave-one-out cross validation. The ex ante approach of the tuning parameters selection is used so that the 2-step Egalitarian LASSO remains comparable with the other presented forecast combination methods. Denoting the pair of shrinkage parameters (λ_1, λ_2) as $\boldsymbol{\lambda}$, the optimal pair is found as follows:

$$\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}} \left(\frac{1}{T} \sum_{t=1}^{T} (y_t - f_{C,t}(\boldsymbol{\lambda}))^2 \right),$$

where

$$f_{C,t}(\boldsymbol{\lambda}) = \hat{\boldsymbol{\beta}}'_{EgalLASSO}(t, \boldsymbol{\lambda}) \boldsymbol{f}_t^*(t, \boldsymbol{\lambda}),$$

where $\boldsymbol{f}_t^*(t, \boldsymbol{\lambda}) = (f_{1,t}^*(t, \boldsymbol{\lambda}), \ldots, f_{k,t}^*(t, \boldsymbol{\lambda}))'$ is the $(k \times 1)$ vector of individual forecasts at time t selected by (3.31) using $\boldsymbol{\lambda}$ and the subset of the training sample, which

includes all but the period t. Similarly, $\hat{\boldsymbol{\beta}}_{EgalLASSO}$ is the $(k \times 1)$ vector of weights obtained from (3.32) using $\boldsymbol{\lambda}$ and the reduced sample. In order to decrease the computational intensity in our empirical applications, we apply the standard 5-fold cross-validation. As in Diebold & Shin (2017), each element of the optimal $\boldsymbol{\lambda}$ is searched trough a grid of plausible values. We find appropriate to use the grid:

$$\left\{\exp\left(-20 + i\frac{22}{19}\right)\right\}_{i=0,\ldots,19}$$

in the forecasting of U.S. Treasury futures volatility application and the grid:

$$\left\{\exp\left(-6 + i\frac{8}{19}\right)\right\}_{i=0,\ldots,19}$$

in the ECB SPF application. Finally, the forecast combination from the complete 2-step Egalitarian LASSO procedure with the ex ante $\boldsymbol{\lambda}$ tuning for the out-of-sample period $T+1$ can be obtained as:

$$f_{C,T+1} = \hat{\boldsymbol{\beta}}'_{EgalLASSO}(\boldsymbol{\lambda}^*)\boldsymbol{f}^*_{T+1}(\boldsymbol{\lambda}^*),$$

where the selection of individual forecasts $\boldsymbol{f}^*_{T+1}(\boldsymbol{\lambda}^*)$ and the weights $\hat{\boldsymbol{\beta}}'_{EgalLASSO}(\boldsymbol{\lambda}^*)$ are obtained using $\boldsymbol{\lambda}^*$ from (3.31) and (3.32) respectively.

3.4 Bayesian Model Averaging Combinations

Bayesian model averaging (BMA) is a technique using bayesian inference to solve the problem of model uncertainty, involving averaging over models including all possible combinations of predictor variables (Raftery *et al.*, 1997). In this thesis, we use bayesian model averaging to combine forecasts in a way inspired by Jacobson & Karlsson (2004) and Eklund & Karlsson (2007), who apply the bayesian model averaging to forecasting Swedish inflation rate. Here we treat the available individual forecasts $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K$ as the set of potential predictors of the variable of interest $\boldsymbol{y}$. We consider a bayesian forecast combination of the forecasts from the set of linear regression models $\mathfrak{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$, where for $j = 1, \ldots, M$, the model $\mathcal{M}_j$ is of the following form:

$$\boldsymbol{y} = \boldsymbol{Z}_j\boldsymbol{\theta}_j + \boldsymbol{\epsilon}, \tag{3.33}$$

where $\boldsymbol{Z}_j = (\mathbf{1}, \boldsymbol{F}_j)$, $\mathbf{1}$ is the $(T \times 1)$ vector of ones and $\boldsymbol{F}_j$ is the $(T \times K_j)$ matrix containing K_j individual forecasts, which are included in the model $\mathcal{M}_j$, as columns. Further, $\boldsymbol{\theta}_j = (\alpha_j, \boldsymbol{\beta}'_j)'$ is the $((1 + K_j) \times 1)$ vector of coefficients and $\boldsymbol{\epsilon}$ is the $(T \times 1)$ vector of errors drawn from $\mathcal{N}(0, \sigma^2\boldsymbol{I})$. The set $\mathfrak{M}$ contains all possible models of this form that the available individual forecasts can give rise to. Because each of the

individual forecasts $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K$ can be either included or excluded from the model, we have in total 2^K different linear regression models to be combined using the bayesian model averaging.

3.4.1 Marginal Likelihood Based Weights

Following Jacobson & Karlsson (2004), we consider the following minimum mean squared error forecast combination at time $T+1$:

$$f_{C,T+1} = \sum_{j=1}^{M} \hat{y}_{j,T+1} p(\mathcal{M}_j|\boldsymbol{y}), \tag{3.34}$$

where $\hat{y}_{j,T+1}$ is the forecast of y_{T+1} obtained from the regression model $\mathcal{M}_j$ (3.33) and the combining weights $p(\mathcal{M}_j|\boldsymbol{y})$ are the posterior probabilities of the respective models. The posterior probabilities are obtained using the Bayes rule:

$$p(\mathcal{M}_j|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{M}_j)p(\mathcal{M}_j)}{\sum_{i=1}^{M} p(\boldsymbol{y}|\mathcal{M}_i)p(\mathcal{M}_i)}, \tag{3.35}$$

where $p(\boldsymbol{y}|\mathcal{M}_j)$ is the marginal likelihood and $p(\mathcal{M}_j)$ is the prior probability of the model $\mathcal{M}_j$. As in Jacobson & Karlsson (2004), we select a diffuse prior for the variance:

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

and for the intercept:

$$p(\alpha_j) \propto 1.$$

For the remaining regression coefficients, we use the g-prior:

$$p(\boldsymbol{\beta}_j|\sigma^2, \mathcal{M}_j) \sim \mathcal{N}\left(0, c\sigma^2(\boldsymbol{F}_j'\boldsymbol{F}_j)^{-1}\right),$$

where we set $c = K^2$. Although the resulting marginal likelihood $m(\boldsymbol{y}|\mathcal{M}_j)$ in this setting is indeterminate, it can be shown that:

$$m(\boldsymbol{y}|\mathcal{M}_j) \propto (c+1)^{-K_j} S_j^{-\frac{T-1}{2}}, \tag{3.36}$$

where

$$S_j = \frac{c}{c+1}(\boldsymbol{y} - \boldsymbol{Z}_j\hat{\theta}_j)'(\boldsymbol{y} - \boldsymbol{Z}_j\hat{\theta}_j) + \frac{1}{c+1}(\boldsymbol{y} - \bar{y}\boldsymbol{1})'(\boldsymbol{y} - \bar{y}\boldsymbol{1}),$$

where $\hat{\theta}_j$ is the OLS estimate of the coefficients from the regression (3.33) and $\bar{y} = 1/T \sum_{t=1}^{T} y_t$. A suggested model prior probability is:

$$p(\mathcal{M}_j) \propto \prod_{i=1}^{K} w_i^{\gamma_{i,j}} (1 - w_i)^{1-\gamma_{i,j}}, \tag{3.37}$$

where w_i is the prior probability of the individual forecast $\boldsymbol{f}_i$ being included in the true model and $\gamma_{i,j}$ is the indicator of whether the individual forecast $\boldsymbol{f}_i$ is in fact included in the model $\mathcal{M}_j$. Following Jacobson & Karlsson (2004), we set $w_i = 1/2$ for $\forall i$, which results in all models from $\mathfrak{M}$ having equal prior probabilities. In summary, the weights for the forecast combination (3.34) (the model posterior probabilities (3.35)) can be computed using the marginal likelihoods (3.36) and the model prior probabilities (3.37).

Note that even though we are primarily combining the linear models based on the individual forecasts rather than individual forecasts themselves, we can rewrite the combination (3.34) in the following way:

$$\begin{aligned}
f_{C,T+1} &= \sum_{j=1}^{M} \hat{y}_{j,T+1} p(\mathcal{M}_j|\boldsymbol{y}) \\
&= \sum_{j=1}^{M} \left(\left(\hat{\alpha}_j + \sum_{\forall i: \boldsymbol{f}_i \in \mathcal{M}_j} f_{i,T+1} \hat{\beta}_i \right) p(\mathcal{M}_j|\boldsymbol{y}) \right) \\
&= \sum_{j=1}^{M} \hat{\alpha}_j p(\mathcal{M}_j|\boldsymbol{y}) + \sum_{j=1}^{M} \left(\sum_{\forall i: \boldsymbol{f}_i \in \mathcal{M}_j} f_{i,T+1} \hat{\beta}_i p(\mathcal{M}_j|\boldsymbol{y}) \right) \\
&= \sum_{j=1}^{M} \hat{\alpha}_j p(\mathcal{M}_j|\boldsymbol{y}) + \sum_{i=1}^{K} \left(f_{i,T+1} \left(\sum_{\forall j: \boldsymbol{f}_i \in \mathcal{M}_j} \hat{\beta}_j p(\mathcal{M}_j|\boldsymbol{y}) \right) \right) \\
&= \omega_0 + \sum_{i=1}^{K} f_{i,T+1} \omega_i.
\end{aligned}$$

Thus it is possible express the resulting forecast combination using the BMA method as a linear combination of the individual forecasts plus a constant.

The disadvantage of the described BMA method is the fast scaling computational intensity. With each additional individual forecast the number of elements of the set $\mathfrak{M}$ and so the number of models for which the posterior probability is to be found increases exponentially. A possible solution is to apply some stochastic search algorithm to efficiently search the model space. For that purpose (similarly to Jacobson & Karlsson (2004)), we use the Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) algorithm proposed by Green (1995). The algorithm is described as follows. In the beginning of each iteration step the chain is at state defined by the

model $\mathcal{M}$. We attempt the move (1) with the probability $p_{(1)}$ or the move (2) with the probability $1 - p_{(1)}$. The moves are:

(1) Draw at random a single individual forecast from the set $\{\boldsymbol{f}_i | i = 1, \ldots, K\}$, each with the probability $1/K$. If the drawn forecast is included in the model $\mathcal{M}$, drop it. If it is not included, add it instead. Propose the newly obtained model $\mathcal{M}^*$ for the acceptance.

(2) Draw at random a single individual forecast from the set $\{\boldsymbol{f}_i | i = 1, \ldots, K \wedge \boldsymbol{f}_i \in \mathcal{M}\}$, each with the probability $1/K_j$, and another individual forecast from the set $\{\boldsymbol{f}_i | i = 1, \ldots, K \wedge \boldsymbol{f}_i \notin \mathcal{M}\}$, each with the probability $1/(K - K_j)$. Swap these two forecasts in the model $\mathcal{M}$. Propose the newly obtained model $\mathcal{M}^*$ for the acceptance.

The proposed model $\mathcal{M}^*$ is accepted with the probability:

$$\alpha = \min\left(1, \frac{m(\boldsymbol{y}|\mathcal{M}^*)}{m(\boldsymbol{y}|\mathcal{M})}\right),$$

which has been simplified from the original acceptance probability in the general algorithm in Green (1995), because in this case the probability $p(\mathcal{M}^*|\mathcal{M}) = p(\mathcal{M}|\mathcal{M}^*)$ and the model prior is uniform. If the model $\mathcal{M}^*$ is accepted, it becomes the new $\mathcal{M}$ for the next iteration step. Otherwise, the model $\mathcal{M}$ is retained (Jacobson & Karlsson, 2004).

Let us define $\mathfrak{M}^*$ the set of models visited by the chain during the iteration process. Although $\mathfrak{M}^*$ can be considerably smaller than $\mathfrak{M}$ in size, it can account for most of the total posterior probability mass, depending on the shape of the posterior (George & McCulloch, 1997). All the calculations of the posterior model probabilities and the resulting forecast combination are then done conditionally on $\mathfrak{M}^*$. In the ECB SPF empirical application, we set the probabilities of both moves types to be equal (i.e. $p_{(1)} = 0.5$) and we run the RJ-MCMC for 6000 iterations, from which the first 1000 are discarded as the burnin. In the U.S. Treasury futures volatility forecasting application, since the amount of individual forecasts and hence the linear regression models is lower, we find it computationally more efficient to compute the posterior probabilities for the entire space of models $\mathfrak{M}$ directly, rather than apply the search via RJ-MCMC.

3.4.2 Predictive Likelihood Based Weights

Eklund & Karlsson (2007) build up on the method by Jacobson & Karlsson (2004) and suggest using the predictive density instead of the marginal likelihood in the calculation of the model posterior probabilities (3.35). They divide the sample into

two parts. First is used to update the prior probabilities on the parameters and second is used to asses the fit of the model. The vector containing the dependent variable $\boldsymbol{y}$ of length $m+l$ is thus divided into the training part $\boldsymbol{y}^*$ of length m and the hold-out part $\tilde{\boldsymbol{y}}$ of length l. For each model $\mathcal{M}_j$ from the set $\mathfrak{M}$, the same applies to the corresponding $((m+l)\times(1+K_j))$ design matrix $\boldsymbol{Z}_j$, which is divided into $(m\times(1+K_j))$ matrix $\boldsymbol{Z}_j^*$ and $(l\times(1+K_j))$ matrix $\tilde{\boldsymbol{Z}}_j$.

Using the same priors as in the previous section, for each posterior predictive density it holds that:

$$p(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{Z}}_j) \propto \frac{(S_j^*)^{\frac{m}{2}}|\boldsymbol{\Lambda}_j^*|^{\frac{1}{2}}}{|\boldsymbol{\Lambda}_j^*+\tilde{\boldsymbol{Z}}_j'\tilde{\boldsymbol{Z}}_j|^{\frac{1}{2}}}\left(S_j^*+(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{Z}}_j\boldsymbol{\gamma}_j)'(\boldsymbol{I}+\tilde{\boldsymbol{Z}}_j(\boldsymbol{\Lambda}_j^*)^{-1}\tilde{\boldsymbol{Z}}_j')^{-1}(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{Z}}_j\boldsymbol{\gamma}_j)\right)^{-T/2}, \tag{3.38}$$

where

$$\boldsymbol{\Lambda}_j^* = \frac{c+1}{c}(\boldsymbol{Z}_j^*)'\boldsymbol{Z}_j^*,$$

$$\boldsymbol{\gamma}_j = \frac{c}{c+1}\hat{\boldsymbol{\theta}}_j^*,$$

$$S_j^* = \frac{c}{c+1}(\boldsymbol{y}^*-\boldsymbol{Z}_j^*\hat{\boldsymbol{\theta}}_j^*)'(\boldsymbol{y}^*-\boldsymbol{Z}_j^*\hat{\boldsymbol{\theta}}_j^*)+\frac{1}{c+1}(\boldsymbol{y}^*-\bar{y}^*\mathbf{1})'(\boldsymbol{y}^*-\bar{y}^*\mathbf{1}),$$

where $\hat{\boldsymbol{\theta}}_j^*$ is the $((1+K_j)\times 1)$ vector of parameters from the regression (3.33) estimated on the training sample and $\bar{y}^* = 1/T\sum_{t=1}^m y_t^*$. As oppose to Jacobson & Karlsson (2004), Eklund & Karlsson (2007) use the parameter c equal to K^3. The model posterior probabilities are then calculated as:

$$p(\mathcal{M}_j|\tilde{\boldsymbol{y}},\boldsymbol{y}^*) = \frac{p(\tilde{\boldsymbol{y}}|\boldsymbol{y}^*,\mathcal{M}_j)p(\mathcal{M}_j)}{\sum_{i=1}^{M}p(\tilde{\boldsymbol{y}}|\boldsymbol{y}^*,\mathcal{M}_i)p(\mathcal{M}_i)}, \tag{3.39}$$

where $p(\tilde{\boldsymbol{y}}|\boldsymbol{y}^*,\mathcal{M}_j)$ is the predictive density corresponding to the model $\mathcal{M}_j$. Finally, the combined forecast is obtained as:

$$f_{C,T+1} = \sum_{j=1}^{M}\hat{y}_{j,T+1}p(\tilde{\boldsymbol{y}}|\boldsymbol{y}^*,\mathcal{M}_j), \tag{3.40}$$

In case the dimensionality of the problem is large, the RJ-MCMC algorithm can be applied again and the posterior probabilities are then calculated conditional on the set of the model visited by the chain $\mathfrak{M}^*$. The RJ-MCMC algorithm used here is the same as described in the previous section, with the exception of the acceptance probability, which is now defined as follows:

$$\alpha = \min\left(1, \frac{p(\tilde{\boldsymbol{y}}|\boldsymbol{y}^*,\mathcal{M}^*)}{p(\tilde{\boldsymbol{y}}|\boldsymbol{y}^*,\mathcal{M})}\right)$$

Eklund & Karlsson (2007) argue that the BMA combination method using weights based on the predictive likelihood instead of the marginal likelihood is less prone to in-sample overfitting, while it still leads to a consistent model selection. The drawback of this method is the trade-off in choosing the length of the training sample and the hold-out sample. By increasing the length of the hold-out sample l, the predictive density becomes more stable. On the other hand, less relevant observations are thus left for updating the prior on parameters. According to empirical findings of Eklund & Karlsson (2007), about 70% of the sample should be hold-out. We find this setting suitable for the macroeconomic SPF ECB application, but we reduce the share of the hold-out sample to 10% for the substantially larger datasets in the forecasting of the U.S. Treasury futures volatility application. Further, we apply the same RJ-MCMC parameter setting as in the marginal likelihood based weights method.

3.5 Alternative Methods

The last section contains forecast combinations via use of ANN's, bagging and boosting. These are frequently used machine learning methods and can be considered alternative ways of combining forecasts. We therefore considered it worthwhile to present it along and compare to the more traditional forecast combining methods.

3.5.1 Artificial Neural Network

The idea of using Artificial Neural Networs (ANNs), a semiparametric modelling technique, for combining forecasts was first presented by Donaldson & Kamstra (1996). The motivation is such that as oppose to the traditional combining methods, ANNs are flexible and can capture even highly nonlinear relationships between the individual forecasts $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K$ and the variable of interest $\boldsymbol{y}$. This includes the potential interactions among the individual forecasts (Donaldson & Kamstra, 1999). Donaldson & Kamstra (1996) employ a single hidden-layer ANN with up to three logistic nodes and the possibility to be complemented with a linear node, formally written as:

$$y_t = \beta_0 + \sum_{k=1}^{n} \beta_k f_{k,t} + \sum_{i=1}^{p} \delta_i \Psi(\boldsymbol{z}_t, \boldsymbol{\gamma}_i) \quad \text{for } t = 1, \ldots, T, \tag{3.41}$$

where we consider:

$$n \in \{0, K\}, \quad p \in \{0, 1, 2, 3\}.$$

The logistic node in (3.41) is defined as:

$$\Psi(\boldsymbol{z}_t, \boldsymbol{\gamma}_i) = \frac{1}{1 + \exp\left(-\left(\gamma_{i,0} + \sum_{k=1}^{K} \gamma_{i,k} z_{k,t}\right)\right)},$$

with the standardized individual forecasts:

$$z_{k,t} = \frac{f_{k,t} - \bar{y}}{s_y},$$

where the sample mean $\bar{y}$ and the sample standard deviation s_y of the variable of interest are computed as:

$$\bar{y} = \frac{1}{T} \sum_{i=1}^{T} y_i,$$

$$s_y = \sqrt{\frac{1}{T-1} \sum_{i=1}^{T} (y_i - \bar{y})^2}.$$

As in Donaldson & Kamstra (1996), the estimation of the ANN is done in the following way. Firstly, denoting the set $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3\}$ as $\boldsymbol{\Gamma}$, 10 $\boldsymbol{\Gamma}$'s are randomly drawn from the multivariate uniform distribution $\mathcal{U}_{(K+1)\times 3}(-1, 1)$. Secondly, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ from the regression (3.41) are estimated by OLS for each drawn $\boldsymbol{\Gamma}$ and the model specification further determined by the number of nodes. Excluding the model with only the intercept, we get 60 different ANNs plus a single fully linear model specification to be estimated. The optimal specification $\{n^*, p^*, \boldsymbol{\Gamma}^*\}$ is then determined in the standard k-fold cross-validation exercise. As in some of the other methods, we use the 5-fold cross-validation. The final ANN forecast combination for the out-of-sample period $T + 1$ can be computed as follows:

$$f_{C,T} = \hat{\beta}_0 + \sum_{k=1}^{n^*} \hat{\beta}_k f_{k,T} + \sum_{i=1}^{p^*} \hat{\delta}_i \Psi(\boldsymbol{z}_T, \boldsymbol{\gamma}_i^*),$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$ are estimates of the parameters from the regression (3.41) using the optimal specification $\{n^*, p^*, \boldsymbol{\Gamma}^*\}$ and the whole sample for the estimation.

3.5.2 Evolving Artificial Neural Network

Much like Donaldson & Kamstra (1996), Harrald & Kamstra (1997) use the ANN to find a nonlinear combination of forecasts of different volatility models. The considered single hidden-layer Evolving Artificial Neural Network (EP-NN) is the following:

$$y_t = \beta_0 + \sum_{k=1}^{K} \beta_k f_{k,t} + \sum_{i=1}^{3} \delta_i \Psi(\boldsymbol{z}_t, \boldsymbol{\gamma}_i) \quad \text{for } t = 1, \ldots, T, \tag{3.42}$$

where again the logistic node is defined as:

$$\Psi(\boldsymbol{z}_t, \boldsymbol{\gamma}_i) = \frac{1}{1 + \exp\left(-\left(\gamma_{i,0} + \sum_{k=1}^{K} \gamma_{i,k} z_{k,t}\right)\right)},$$

Newly, Harrald & Kamstra (1997) utilize the means of evolutionary programming and search for the optimal set of parameters $\boldsymbol{\Gamma}^* = \{\boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*, \boldsymbol{\gamma}_3^*\}$ from (3.42) using the stochastic numerical process. The algorithm for the estimation of their EP-NN has the following steps:

1) For each parent $p \in \{1, \ldots, n\}$, make an independent random draw $\boldsymbol{\Gamma}_p$ from the multivariate uniform distribution $\mathcal{U}_{(K+1)\times 3}(-1, 1)$

2) For each $p \in \{1, \ldots, n\}$, estimate the model (3.42) by OLS using $\boldsymbol{\Gamma}_p$.

3) Sort the vector $(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_n)$ by the in-sample MSE produced by respective models from the previous step.

4) For each $p > n/2$, replace $\boldsymbol{\Gamma}_p$ by $\boldsymbol{\Gamma}_p + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathcal{N}_{(K+1)\times 3}(0, \sigma)$

5) Repeat steps 2)-4) for g generations.

6) Choose $\boldsymbol{\Gamma}_1$, the set of parameters from the model with the lowest MSE overall, as the optimal set $\boldsymbol{\Gamma}^*$.

Additionally, as a protective measure against overfitting, Harrald & Kamstra (1997) propose to run the whole procedure independently 29 times and select the EP-NN with the median MSE as the final model. We run the procedure only once in order to accelerate the computations. The results in our empirical applications are reported for the parameter σ equal to 0.05. Further, we use the size of the population $n = 16$ and evolve the network for $g = 200$ generations.

3.5.3 Bagging

Bootstrap Aggregation (Bagging) is a method designed for situations, when it is necessary to deal with a big amount of potential predictors and unstable model decision rules. Assuming the covariance stationary environment, the use of bagging is expected to reduce the out-of-sample mean squared forecast error (Inoue & Kilian, 2008). In this thesis, we consider the bagging technique applied to an OLS regression with a pre-test, as it is done by Inoue & Kilian (2008), a self-contained forecast combination method.

Let's consider the following unrestricted regression model:

$$y_t = \boldsymbol{f}_t'\boldsymbol{\beta} + \epsilon_t \quad \text{for } t = 1, \ldots, T, \tag{3.43}$$

where y_t is the forecast of the variable of interest at time t, $\boldsymbol{f}_t$ is the vector of K individual forecasts, $\boldsymbol{\beta}$ is the vector of parameters and ϵ_t is the error term. We can construct and estimate the following pre-test model using the two sided t-test:

$$\boldsymbol{y}_b = \boldsymbol{\mathcal{F}}_b \boldsymbol{\gamma}_b + \boldsymbol{\nu}, \tag{3.44}$$

where $\boldsymbol{\mathcal{F}}_b$ is the matrix of individual forecasts, containing only those individual forecasts $\boldsymbol{f}_j$ $(j = 1, \ldots, K)$ for which $|t_j| > 1.96$. Where t_j denotes the t-statistic for the null hypothesis that β_j from the unrestricted model (3.43) equals zero.

Further, consider the set of bootstrap samples $\{(\boldsymbol{y}_b^*, \boldsymbol{F}_b^*)\}_{b=1,\ldots,B}$, where the individual samples are compound of p blocks of size m drawn randomly with replacement from the following matrix:

$$\begin{pmatrix} y_1 & f_{1,1} & f_{2,1} & \ldots & f_{K,1} \\ y_2 & f_{1,2} & f_{2,2} & \ldots & f_{K,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_T & f_{1,T} & f_{2,T} & \ldots & f_{K,T} \end{pmatrix}.$$

Noting that $p = \lfloor T/m \rfloor$. This moving block bootstrap is used to retain the serial dependency of the original series in the bootstrap samples (Gonçalves & White, 2004). The block size m was chosen to be equal to $\lfloor \sqrt[3]{T} \rfloor$ for simplicity. For the discussion of admissible block sizes see e.g. Politis *et al.* (1997). For each bootstrap sample, we first estimate the unrestricted model and obtain $\hat{\boldsymbol{\beta}}_b^*$. Then, we construct and estimate the pre-test model and obtain $\hat{\boldsymbol{\gamma}}_b^*$. The t-statistic for pre-testing $|t_j^*|$ is computed as follows (Inoue & Kilian, 2008):

$$|t_j^*| = \left| \frac{\hat{\beta}_{j,b}^*}{\sqrt{\mathrm{Var}(\hat{\beta}_{j,b}^*)}} \right|,$$

where

$$\mathrm{Var}(\hat{\beta}_{j,b}^*) = \frac{1}{\sqrt{T}} \left((\hat{\boldsymbol{H}}^*)^{-1} \hat{\boldsymbol{S}}^* (\hat{\boldsymbol{H}}^*)^{-1} \right)_{jj}$$

and where

$$\hat{\boldsymbol{S}}^* = \frac{1}{pm} \sum_{k=1}^{p} \sum_{i=1}^{m} \sum_{j=1}^{m} (\boldsymbol{f}_{(k-1)m+i}^* \epsilon_{(k-1)m+i}^*)(\boldsymbol{f}_{(k-1)m+j}^* \epsilon_{(k-1)m+j}^*)',$$

$$\hat{\boldsymbol{H}}^* = \frac{1}{pm} \sum_{k=1}^{p} \sum_{i=1}^{m} (\boldsymbol{f}_{(k-1)m+i}^* (\boldsymbol{f}_{(k-1)m+i}^*)'),$$

where $\epsilon_t^* = y_t^* - (\boldsymbol{f}_t^*)' \hat{\boldsymbol{\beta}}$. The bagging forecast combination at the out-of-sample

period $T+1$ is then computed as:

$$f_{C,T+1} = \frac{1}{B}\sum_{b=1}^{B}(\boldsymbol{f}_{t+1})'\hat{\boldsymbol{\gamma}}_b.$$

In both of our applications, we use the number bootstrap replications B equal to 500. Further, we use the classical 1.96 threshold for the t-statistic as suggested by Inoue & Kilian (2008) for the ECB SPF application. Nevertheless, we reduce the threshold to 1.28 for the forecasting of U.S. Treasury futures realized volatility application, because we find the original threshold too strict for pre-testing in environments with strongly correlated forecasts.

3.5.4 Componentwise Boosting

Boosting is a functional gradient descent technique, which originated from the machine learning community and was firstly applied to classification problems (Bühlmann & Yu, 2003). It is generally suitable for cases where there are high number of potential predictors and some form of information condensation or variable selection is required (Buchen & Wohlrabe, 2011). The componentwise L_2-boosting was firstly presented by Bühlmann & Yu (2003). Here we apply the algorithm as described by Buchen & Wohlrabe (2011) to the following linear regression forecast combining model:

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{Z},\boldsymbol{\beta}) = \boldsymbol{Z}\boldsymbol{\beta} =: \boldsymbol{\Psi}(\boldsymbol{Z},\boldsymbol{\beta}), \tag{3.45}$$

where $\boldsymbol{y}$ is the $(T\times 1)$ vector of dependent variable, $\boldsymbol{Z} = (\mathbf{1}, \boldsymbol{F})$ is the $(T\times(K+1))$ design matrix and $\boldsymbol{F} = (\boldsymbol{f}_1,\ldots,\boldsymbol{f}_K)$ is the $(T\times K)$ matrix of individual forecasts. The coefficients corresponding to the individual forecasts, which are not selected by the boosting algorithm are restricted to be zero. We use OLS as the base learner $\boldsymbol{\psi}(.)$ and the quadratic loss function:

$$L(\boldsymbol{y},\Psi(\boldsymbol{Z},\boldsymbol{\beta})) = \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\Psi}(\boldsymbol{Z},\boldsymbol{\beta}))^2$$

The steps of the algorithm are following:

1) Set $m=0$. Initialize $\hat{\boldsymbol{\psi}}_0 = \frac{1}{T}\mathbf{1}'\boldsymbol{y}$.

2) Increase m by 1. Compute the negative gradient vector $-\frac{\partial L(\boldsymbol{y},\Psi)}{\partial\Psi}$ evaluated at $\hat{\boldsymbol{\psi}}_{m-1}(\boldsymbol{Z},\hat{\boldsymbol{\beta}}^{[m-1]})$: $\boldsymbol{u} = \boldsymbol{y} - \hat{\boldsymbol{\psi}}_{m-1}(\boldsymbol{Z},\hat{\boldsymbol{\beta}}^{[m-1]})$.

3) For $k=1,\ldots,K$ regress $\boldsymbol{u}$ on $\boldsymbol{f}_k$ and compute the sum of squared residuals $SSR_k = \mathbf{1}'(\boldsymbol{u}-\hat{\theta}_k\boldsymbol{f}_k)^2$.

4) Find $k^* = \arg\min_k SSR_k$.

5) Update $\hat{\boldsymbol{\psi}}_m(\boldsymbol{Z}, \hat{\boldsymbol{\beta}}^{[m]}) = \hat{\boldsymbol{\psi}}_{m-1}(\boldsymbol{Z}, \hat{\boldsymbol{\beta}}^{[m-1]}) + \nu\hat{\theta}_{k^*}\boldsymbol{f}_{k^*}$.

6) Repeat steps 2-5 until $m = M$.

The shrinkage parameter ν was set equal to 0.1. It was found that this boosting algorithm is not very sensitive to the choice of ν (Buchen & Wohlrabe, 2011). The number of boosting iterations M was determined in a standard 5-fold cross-validation exercise. The function $\boldsymbol{\Psi}(\boldsymbol{Z}, \boldsymbol{\beta})$ is finally estimated as:

$$\hat{\boldsymbol{\Psi}}(\boldsymbol{Z}, \hat{\boldsymbol{\beta}}^{[M]}) = \hat{\boldsymbol{\psi}}_m(\boldsymbol{Z}, \hat{\boldsymbol{\beta}}^{[M]}).$$

The componentwise boosting forecast combination for the out-of-sample period $T+1$ is then computed as:

$$f_{C,T+1} = \hat{\Psi}(\boldsymbol{z}_{T+1}, \hat{\boldsymbol{\beta}}^{[M]}),$$

where $\boldsymbol{z}_{T+1} = (1, f_{1,T+1}, \ldots, f_{K,T+1})'$.

3.5.5 AdaBoost

The AdaBoost is a commonly applied Boosting Algorithm (Barrow & Crone, 2016). Barrow & Crone (2016) empirically test different choices of AdaBoost metaparameters on time series data. They introduce the AdaBoost.BC method, which combines the metaparameters that came out best in the test. Here we follow their methodology and use AdaBoost.BC to combine individual forecasts.

In the beginning of each iteration step of the AdaBoost algorithm, we sample a training set of length size T from the original sample with replacement. In the $i-th$ step, for $t = 1, \ldots, T$, the vector $(y_t, f_{1,t}, \ldots, f_{K,t})$ has the probability $p_t^i = w_t^i / \sum_{t=1}^{T} w_t^i$ of being drawn. The initial weights w_t^1 are set equal to 1 for all observations, thus inducing a uniform distribution. Then we train the base learner, which is in this case the multilayer perceptron (MLP). The MLP has one hidden layer and two hidden nodes. We use the hyperbolic tangent activation function for the hidden nodes and linear activation function for the output node. Further, we use the L-BFGS, an optimizer from the class of quasi-Newton methods, to optimize the weights in the MLP. For details on MLP's refer to Zhang *et al.* (1998) and Demuth *et al.* (2014).

Using the OLS, the parameters of the following regression model are then estimated on the training sample:

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.46}$$

where $\boldsymbol{y}$ is the dependent variable of interest, $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K)$ is the matrix of individual forecasts, $\boldsymbol{\beta}$ is the vector of parameters and $\boldsymbol{\epsilon}$ is the vector of errors.

Next, the threshold-based loss L_t^i is calculated for each observation:

$$L_t^i = \begin{cases} 1, & \text{if } ARE_t^i > \phi, \\ 0, & \text{else,} \end{cases}$$

where the absolute relative error is computed as:

$$ARE_t^i = \left| \frac{y_t - \hat{y}_t^i}{y_t} \right|,$$

where $\hat{y}_t^i$ is the prediction produced from the trained MLP. Then the weighted average loss is computed as:

$$\bar{L}^i = \sum_{t=1}^{T} p_t^i L_t^i.$$

We use the threshold parameter ϕ equal to 0.1. Next, we calculate the model confidence measure:

$$\beta_i = \log\left(\frac{1}{\bar{L}^i}\right)$$

The observation weights are then updated according to the following rule:

$$w_t^{i+1} = w_t^i \beta_i^{(1-L_t^i)}.$$

The algorithm is run for 50 iterations. Finally, the forecast combination at time $T+1$ is computed by simply averaging over the predictions produced from the MLP's trained at different iterations:

$$f_{C,T+1} = \frac{1}{50} \sum_{i=1}^{50} \hat{y}_{T+1}^i.$$

This last step is the inspiration from the research on forecasts combinations (Barrow & Crone, 2016).

3.6 Artificial Prediction Markets

From the various types and modifications of the artificial prediction market methods summarized in the literature review 2.6, we apply in our applications the two variants of continuous Artifical Prediction Markets, because they are readily usable for the combining of the time series forecast data, unlike most of the rest of the artificial prediction market methodology, which is dedicated to the classification problems only. Also, we present the original Market for Kernels method, which is our modest attempt to extend artificial prediction market literature for a method directly applicable to time series regression problems.

3.6.1 Continuous Artificial Prediction Markets

The continuous Artificial Prediction Markets (c-APM) is the prediction or in our case forecast combination method designed by Jahedpari *et al.* (2017) for regression problems. The c-APM agents participate in the artificial prediction markets with the parimutuel betting mechanism and a market maker. Jahedpari *et al.* (2017) suggest two variants of their method: c-APM with agents having constant betting functions, c-APM with agents who observe predictions of other agents and adapt their strategies via the reinforcement technique called Q-learning. The c-APM training algorithm is following (Jahedpari *et al.*, 2017):

1) Initialize agents with equal budgets.

2) Initialize the market for the prediction of the variable y_t. Repeat the steps 2 – 11 for $t = 1, \ldots, T$.

3) In the first round of the market, every agent bets a $MaxRPT$ (maximum rate per transaction) share of her budget on her own individual prediction (forecast).

4) For all remaining rounds of the market, repeat the steps 5) and 6).

5) Each agent decides about her prediction and how much she bets on it, depending on the type of the agent. In the c-APM (Constant), the agents with constant betting functions simply bet a fixed share ($MaxRPT$) of their budget on their own individual prediction in every round. Whereas in the c-APM (Q-learning), the agents observe the market prediction announced by the market maker at the end of each round and have two options. They can either preserve their current predictions or adjust them. Firstly, each agent estimates the error of her current prediction:

$$estError_k = Prediction - prediction_k,$$

where $Prediction$ is the market prediction from the previous round. Based on the estimated error, each agents identifies the state s she is in, which is defined by the current round number and the cluster into which the error falls (small, medium, large). The agents then have the option to change their predictions in a following way:

$$prediction_k \leftarrow prediction_k + \delta_{k,s} \times estError_k,$$

where $\delta_{k,s}$ is the parameter from the interval $[0, 1]$ reflecting the confidence in the wisdom of the crowd of the k-th agent in the state s. The agents always choose the action ("preserve" or "change"), which has the higher Q-value for the given agent and state. Further, the agents decide how much to bet. They estimate the

scores of their newly chosen predictions:

$$score'_k = \log(accuracy'_k),$$

where

$$accuracy'_k = \max\left\{100\left(1 - \frac{|Prediction - prediction_k|}{oet}\right), 1\right\},$$

where oet is the outlier error threshold from the previous market. If the $score'_k$ of the k-th agent is greater than or equal to 1, the agent expects that a bet at this prediction will result in her winning money and is thus motivated to bet as much as possible ($MaxRPT$). In the opposite case of $score'_k < 1$, the agent expects to lose money and thus bets as little as possible ($MinRPT$).

6) At the end of the round, the market maker aggregates individual predictions of all agents into a single market prediction:

$$Prediction = \frac{\sum_{k=1}^{K} bet_k \times prediction_k}{\sum_{k=1}^{K} bet_k},$$

where K is the number of agents in the market. Sizes of bets represent natural weights of predictions as they reflect the confidence of respective agents in their own predictions. Also, this weighting scheme promotes the principle, that agents who are often more accurate in predictions and so accumulate greater budgets, have greater impact on the aggregated market prediction.

7) At the end of the market, the true outcome y_t is revealed and the outlier error threshold (eot) for the current market is computed using the inter-quartile range.

8) The error clusters are recomputed using the online k-means clustering algorithm (MacQueen *et al.*, 1967).

9) Agents obtain revenue for each of their bet according to the formula:

$$revenue = score \times bet,$$

where

$$score = \log(accuracy),$$

and

$$accuracy = \max\left\{100\left(1 - \frac{|y_t - prediction|}{oet}\right), 1\right\}.$$

10) Agents calculate the potential revenue (*potRevenue*) they could have earned for each action in each state they have visited in the market and update the respective Q-values in a following way:

$$Q_{k,s} \leftarrow (1-\alpha)Q_{k,s} + \alpha \times potRevenue,$$

where α is the learning rate parameter from the interval $[0, 1]$.

11) Agents update their confidence in the wisdom of the crowd parameter for each state the have visited in the market:

$$\delta_{k,s} \leftarrow (1-\alpha)\delta_{k,s} + \alpha \times truncate\left(\frac{y_t - prediction_{k,s}}{Prediction_s - prediction_{k,s}}\right),$$

where

$$truncate(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x > 1, \\ x, & \text{else.} \end{cases}$$

The parameter setting used is similar to Jahedpari *et al.* (2017). We set $MaxRPT = 0.9$ for the first round, so the influence the agent's original prediction on her budget remains strong. For all the other round, we set $MaxRPT = 0.01$ and $MinRPT = 0.0001$. Each market is run for 10 rounds and the learning parameter α is set equal to 0.7. The market predictions from the last round of each market are the c-APM predictions. The fully trained artificial prediction market, in terms of the budget distribution, updated Q-values and δ parameters, then can be used to combine forecasts out-of-sample at time $T + 1$ by simply running the market for y_{T+1} without further updates.

3.6.2 Market for Kernels

When designing an artificial prediction market mechanism applicable to a forecast combination or generally a regression problem, one needs somehow overcome the fundamental problem of transferring either from a countable number of outcomes on which the agents can bet to an uncountable number of outcomes, when building up on the idea of Lay & Barbu (2010), or similarly transferring from a countable number of futures to an uncountable number of futures, when following along the lines of Storkey (2011), Millin *et al.* (2012) and Hu & Storkey (2014). Our approach is inspired by Lay & Barbu (2012), who use density estimates of individual agents and a reward kernel to reward the agents at the end of the market based on how close their prediction are from the true values. However, unlike the approach of Lay & Barbu (2012), ours does not require the complete density estimates from the

individual agents directly and so is suitable also for combining point forecasts. We borrow the idea of using kernels in artificial prediction markets from Lay & Barbu (2012), but we use them in a different way than reward kernels. We let the agents bet in each market on the probability density represented by a predefined kernel shifted and scaled according to their choice (hence the term Market for Kernels), which approximately reflects their beliefs about the probability of all outcomes across the outcome space. Then, after the true outcome is revealed, we reward the agents based on the relative values these densities assign to the true outcome and the relative sizes of their bets. The choice of each agent's kernel scale and shift is driven by her expected outcome in the current market and her expected accuracy of her prediction based on the experience from the previous markets.

The Market for Kernels is trained using the following algorithm:

1) Firstly, all of the K agents, corresponding to the K individual forecasts, are initialized with equal budgets $w_i = 1/K$ for $\forall i = 1, \ldots, K$.

2) For each observation at time t in the training sample $t = 1, \ldots, T$, the single round market is run by repeating steps 3,4,5.

3) The market allows one dimensional gaussian kernels defined as:

$$K(x, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right), \tag{3.47}$$

where σ denotes the scale parameter. Each agent specifies the the expected outcome $f_{i,t}$ (prediction), which acts as a shift of the gaussian kernel (3.47) in the outcome space. Further, each agent specifies the kernel scale parameter σ as the average of her absolute prediction errors from the previous markets:

$$\sigma_{i,t} = \frac{1}{t-1} \sum_{\tau=1}^{t-1} |\sigma_{i,t}|,$$

assuming the initialization:

$$\sigma_{i,1} = 1, \quad \text{for } \forall i$$

in the first market. The selected scale reflects the uncertainty of each agent about her own prediction. Regarding the size of the bet, we assume that the agents bet their whole budgets in each market. So, it holds that:

$$bet_{i,t} = budget_{i,t}, \quad \text{for } \forall i, t.$$

4) The market maker aggregates the pool of predictions into a single market predic-

tion using the following formula (as in Jahedpari *et al.* (2017)):

$$f_{C,t} = \frac{\sum_{i=1}^{K} bet_{i,t} \times f_{i,t}}{\sum_{i=1}^{K} bet_{i,t}}.$$

5) The true outcome y_t is revealed and the agents are rewarded according to the accuracy of their predictions and sizes of their bets. The following reward formula is used:

$$reward_{j,t} = \frac{bet_{j,t} \times K_j(y_t - f_{j,t})}{\sum_{i=1}^{K} bet_{j,t} \times K_i(y_t - f_{i,t})}$$

The reward function can be in fact understood as a contributing share of the j-th agent to the market weighted kernel density estimate at the true outcome y_t. Since the agents bet their whole budgets in every market, current rewards of all agents directly translate into their next market budget:

$$budget_{i,t+1} = reward_{i,t}, \quad \text{for } \forall i, t.$$

After the market is fully trained in a sense of the total budget distribution and the updated kernel scales of individual agents, it can be run on the testing data and the obtained aggregate market predictions in the step 4 is taken as the resulting Market for Kernels prediction. The property of infinite support of the gaussian kernel in combination with our reward function ensure that no agent can ever go fully bankrupt, which protects the market from a partial information loss. The design of the market allows each agent to effectively accumulate wealth and thus increase its influence on the aggregate market prediction after a series of accurate predictions from an arbitrary starting position.

We deliberately eliminate the possibility for user specified input parameters. We aim for the method as simple and working as autonomously as possible, while still keeping some degree of flexibility thanks to the elements of learning and combining based on the past accuracy. Although, we acknowledge that the Market for Kernels method is directly extensible by e.g. setting some parameter reducing the window of last observations the agents take into account when deciding about the scale of their kernels, such as the parameter ν in Bates-Granger methods (3.4), (3.5) and (3.6), or perhaps by limiting the share of the budget that the agents can bet in each market by some other parameter, such as $MaxRPT$ in the c-APM presented in the previous section 3.6.1. Furthermore, please note that our primary goal in this study, regarding the Market for Kernels, is to asses its performance empirically against other existing forecast combination methods. We recognise that the Market for Kernels method,

as presented at this stage, is a mere suggestion of an algorithm, inspired by our own experience with combining forecasts, absent of any formal proof of the market convergence towards the theoretically optimal true weights. This work is yet to be carried out in the future.

Chapter 4

Applications

In this thesis, we apply the described forecast combination methods in two distinct application. The time series used in these applications differ completely in their nature and size, which allows us to make a broader assessment of the performance of the individual methods. In the following two sections are presented both the macroeconomic and the financial application.

4.1 ECB Survey of Professional Forecasters

Since the January of 1999, the European Central Bank (ECB) runs the quarterly Survey of Professional Forecasters (SPF) about the anticipation of the future growth of real gross domestic product, inflation and unemployment in the euro area. The results of the survey are published periodically in the ECB Monthly Bulletin. Among the contributors to the survey are relevant professionals affiliated with both financial and non-financial institutions from the European Union (EU) (ECB, 2018). Apart from the forecasts themselves, the SPF contains information about the level of uncertainty of individual forecasts and the mean forecast for each of the macroeconomic variable of interest. The contributing professionals are asked for forecasts at multiple horizons. The ECB SPF data together with the list of contributors is publicly available[1].

Accurate forecasts of the key macroeconomic variables such as those in the SPF are of great value to all sorts of economic agents including policy makers, investors and households. The individual SPF forecasts may differ substantially depending on the information set and the methodology used of the particular contributor. Therefore, naturally, a research has been done on how to optimally combine these forecasts in order to improve accuracy and more specifically, whether there is a combining strategy that could consistently out-perform the simple average of these individual

[1]http://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html

forecasts (e.g. Genre *et al.* (2013), Diebold & Shin (2017) and Conflitti *et al.* (2015). Our motivation in this is application is to tackle the problem of combining forecasts from SPF with a broad range of methods, some of which are rarely seen in economic applications and to our knowledge were not previously applied to this problem, and asses whether any new insights can be drawn. We aim to discuss our experience with combining SPF forecasts in contrast with the up to date findings of other researchers.

4.1.1 Data

In this study, we focus on forecasts of the yearly percentage change of real gross domestic product (RGDP) as according to the definition of European system of national and regional accounts (ESA), yearly percentage change of the Harmonised Index of Consumer Prices as published by Eurostat (HICP, harmonised inflation) and unemployment rate (UNEM) as calculated by Eurostat over 1-year and 2-year horizons. All the considered variables are expressed in percentage points. For a complete and thorough description of the ECB's SPF including details about the history and design of the questionnaire you may refer to Garcia (2003) and Bowles *et al.* (2007). We obviously only work with forecasts with up to the current date horizons, so their actual accuracy can be assessed. The forecasts from the SPF for different macroeconomic variables target different months, nevertheless all the forecast series are quarterly. The real GDP growth forecast series is composed of 73 observation from July 1999 to July 2017 for the 1 year forecast horizon and of 69 observations from July 2000 to July 2017 for the 2 years forecast horizon. The harmonised inflation forecast series is composed of 73 from December 1999 to December 2017 for the 1 year forecast horizon and of 69 observations from December 2000 to December 2017 for the 2 years forecast horizon. The unemployment forecast series is composed of 73 from November 1999 to November 2017 for the 1 year forecast horizon and of 69 observations from November 2000 to November 2017 for the 2 years forecast horizon. The table 4.1 summarizes the descriptive statistics of the SPF target macroeconomic variables series. Note that the 1 year and 2 year forecast horizon series for respective variables differ only by the first year of the survey (1999), which is not present in the 2 year forecast horizon series. From the three examined variables, the unemployment rate has the least variance in relative to its mean, which suggests it might be easiest one to predict. In contrast, the real GDP growth has a variance relative to its mean. Most of it is induced by the global financial crisis slump, which is apparent from the figure 4.1.

Table 4.1: Descriptive statistics of the SPF target macroeconomic variables for the euro area

Statistic	RGDP		HICP		UNEM	
	1Y	2Y	1Y	2Y	1Y	2Y
Mean	1.16	1.05	1.77	1.74	9.41	9.42
Median	1.50	1.40	2.00	2.00	9.10	9.00
Mode	1.60	1.60	2.10	2.10	8.30	8.30
Std. Dev.	1.56	1.53	0.99	1.01	1.39	1.43
Variance	2.44	2.34	0.98	1.01	1.93	2.04
Minimum	-4.80	-4.80	-0.30	-0.30	6.90	6.90
Maximum	3.70	3.40	4.00	4.00	12.20	12.20
Kurtosis	2.94	3.17	-0.38	-0.47	-0.72	-0.84
Skewness	-1.42	-1.53	-0.43	-0.37	0.29	0.27

4.1.2 Individual Forecasts

An important feature of the SPF dataset to mention is that its individual contributors often enter and exit the survey since its beginning. Moreover, the contributors occasionally provide the forecasts for only some of the variables used in this study in a given questionnaire. These facts leave us with unbalanced panels of data to be dealt with. Because majority of the forecast combining methods presented in the chapter 3 require balanced panels, with the exception being e.g. the PEW 3.1.7 of Capistrán & Timmermann (2009), it necessary to pre-process the data. In balancing the panels, we follow the process of Genre *et al.* (2013). Firstly, only the forecasters with no more than 4 consecutive missing observations are kept. Secondly, the missing observations are imputed using the linear filter. The parameter β which defines the filter is obtained by estimation of the following model by OLS:

$$\delta_t = \beta\delta_{t-1} + \epsilon_t, \quad \text{for } \forall t, \tag{4.1}$$

where δ_t is the deviation from the mean forecast of all forecasters at time t. The final number of individual forecasters in the balanced datasets is The estimated beta is then used to impute forward the missing observations and, for simplification, also the missing observations in the very beginning of the sample (in case there are any).

The performance of the individual SPF contributors is summarized in the table 4.2 using the measures defined in the section 5.1. The first observation we can make from the table is that the performance spread between the best and the worst individual is quite small across the measures and variables. This indicates that the forecaster use similar methodologies and make their forecasting decisions based on similar information sets. This point makes it clear why it is challenging for any forecast combination method to significantly out-perform the simple average. Secondly, the mean average percentage errors (MAPEs) are substantially lower for the unemployment rate than the real GDP growth and the harmonised inflation series. This again shows that it might be relatively harder to forecast inflation and GDP growth than the unemployment rate. Thirdly, regarding all the studied measures a variables, the forecast performance is relatively worse for the 2 year forecast horizon than for the 1 year forecast horizon. The natural implication is that forecasting these variables is relatively harder for longer horizons. These findings are supported by the figure 4.1. The figure shows, for example, that the real GDP growth slump in 2009 was not at all anticipated in the 2007 SPF, while it was much more anticipated in the 2008 SPF.

Table 4.2: Forecast performance (measured in terms of RMSE, MAE and MAPE) of indindividual forecasters from the ECB SPF for the target macroeconomic variables and horizons

Measure	Performance	RGDP		HICP		UNEM	
		1Y	2Y	1Y	2Y	1Y	2Y
RMSE	Mean	1.19	1.88	0.94	1.03	0.71	1.36
	Best Individual	1.07	1.77	0.86	0.97	0.66	1.26
	0.25 Quantile	1.15	1.84	0.91	1.01	0.68	1.29
	Median Individual	1.16	1.88	0.94	1.02	0.70	1.33
	0.75 Quantile	1.23	1.91	0.95	1.05	0.73	1.41
	Worst Individual	1.35	2.02	1.05	1.17	0.81	1.55
MAE	Mean	0.86	1.31	0.76	0.82	0.55	1.06
	Best Individual	0.78	1.22	0.67	0.75	0.50	0.92
	0.25 Quantile	0.84	1.27	0.73	0.79	0.54	1.02
	Median Individual	0.85	1.29	0.76	0.81	0.56	1.05
	0.75 Quantile	0.88	1.34	0.78	0.84	0.57	1.11
	Worst Individual	0.96	1.43	0.83	0.96	0.62	1.25
MAPE	Mean	82.95	145.21	172.09	219.56	5.86	11.00
	Best Individual	67.95	129.07	144.44	189.86	5.22	9.53
	0.25 Quantile	78.19	138.68	164.09	208.20	5.77	10.64
	Median Individual	81.74	143.34	172.60	220.05	5.94	10.94
	0.75 Quantile	87.32	152.86	176.43	228.77	6.03	11.46
	Worst Individual	98.78	160.69	210.64	261.36	6.49	12.81
Number of Forecasters		21	19	19	20	17	19

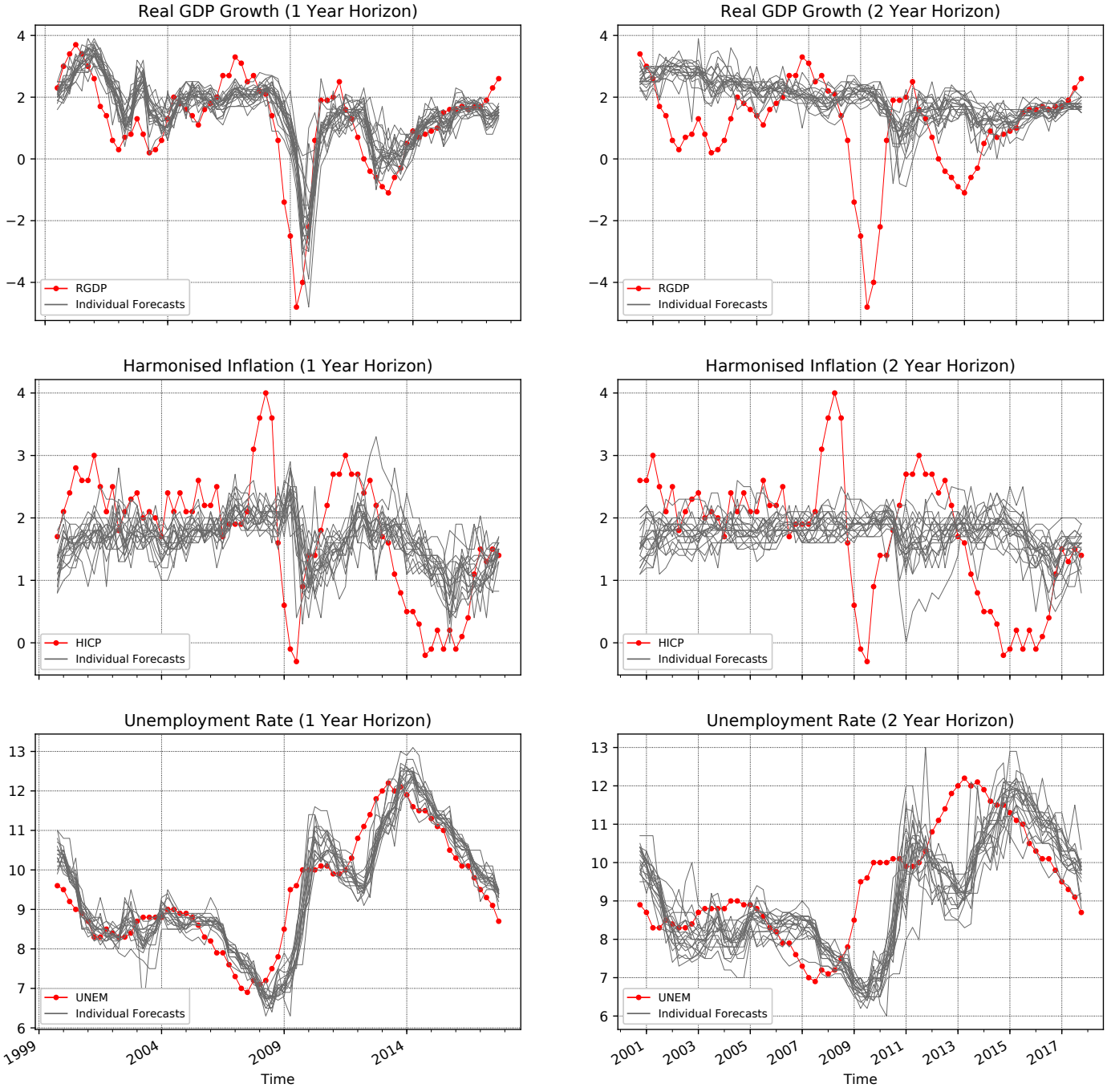

Figure 4.1: Individual forecasts of the macroeconomic variables (in percentage points) from the ECB SPF and the target variables in time

4.2 Forecasting U.S. Treasury Futures Volatility

Volatility of financial assets is a measure, which is of great concern to risk analysts, portfolio analysts as well as academic researchers. An accurate forecast of volatility of an asset over a considered holding period is a base for making an investment decision. Moreover, volatility is the main input in derivative security pricing models (Poon & Granger, 2003). Therefore, a range of econometric methods were developed for modelling and forecasting volatility. Here, we estimate several of the most known volatility models and use it to forecast volatility of log returns of the U.S. Treasury futures. We test whether any enhancement in the volatility forecasts can be obtained by combining the forecast from the individual volatility models.

4.2.1 Data

The U.S. Treasury notes and bonds are securities representing a loan to the U.S. government and providing their holder with semi-annual payments until maturity. In this thesis, we work with U.S. Treasury futures, which are the securities underlay by the U.S. Treasury notes and bonds. Generally, futures contracts are agreements to sell or buy a given security on a specified future date and at a specified price. The U.S. Treasury futures are liquid assets, which offer an opportunity to speculate on interest rates or hedge against risks. In Q1 2018, the U.S. Treasury futures were among top 10 most liquid futures contracts traded at the CME Group, which is the leading market for derivatives (“Most Traded Futures", 2018). More specifically, we use futures for the 2-Year, 5-Year and 10-Year Treasury Notes and the U.S. Treasury Bond with the 30-Year maturity. These are marked with the respective tickers: US, FV, TY and US. Our sample covers the period from the 1st of July 2003 to the 29th of December 2017, which makes for 3635 daily observations. We can therefore consider it a representative sample for the noted futures.

Firstly, in the figure 4.2 are depicted daily log-returns (log-differences of daily closing prices) of the U.S. Treasury futures which, as opposed to the futures prices, are stationary. The results of the Augmented Dickey-Fuller test (Said & Dickey, 1984) are presented in the table 4.3. The results clearly show we can reject the null hypothesis of a unit-root presence for all the log-return series. The descriptive statistics of the log-returns are summarized in the table 4.4. We can confirm that U.S. Treasury futures log-returns show some of the stylized empirical facts for financial assets (Cont, 2001). Namely, the positive excess kurtosis signalizes a leptokurtic underlying distribution with heavy tails for all the studies futures. Moreover, the significant difference between the TU (2 Year) and US (30 Year) futures implies that futures underlay by long term bonds are less prone to extreme price movements in

comparison to futures underlay by short term notes. Also, there are clearly noticeable volatility clusters throughout the examined period in the figure 4.2.

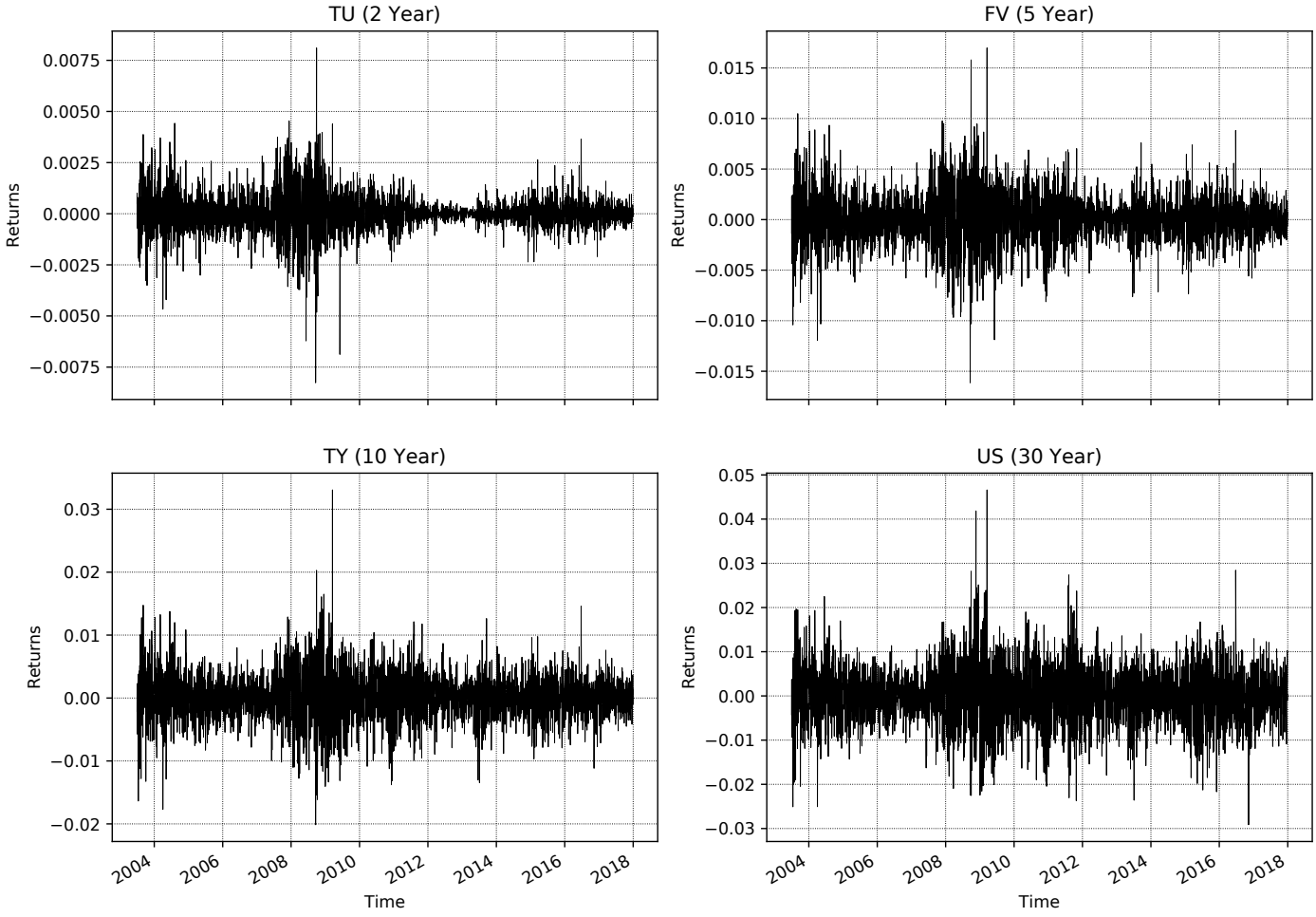

Figure 4.2: Log-returns of U.S. Treasury futures

Our main variable of interest in this application is the volatility of the U.S. Treasury futures log-returns. The problem with modelling volatility is that it is in its nature a latent, unobservable variable. So, in order to be able to assess accuracy, there is a necessity to choose a suitable volatility proxy. Let us assume the following diffuse process:

$$dp(t) = \mu(t)dt + \sigma(t)dW(t), \tag{4.2}$$

where $p(t)$ is the logarithm of the price at time t, $W(t)$ is a Wiener process, $\mu(t)$ is a finite variation process and $\sigma(t)$ is a stochastic process, which is independent of $W(t)$ (Corsi, 2009). The one day integrated variance of such process is then defined as:

$$IV_t = \int_{t-1}^{t} \sigma^2(x)dx, \tag{4.3}$$

Table 4.3: Augmented Dickey-Fuller test results for the log-returns of U.S. Treasury futures

	TU (2 Year)	FV (5 Year)	TY (10 Year)	US (30 Year)
Test Statistic	-22.7751	-35.4464	-12.5631	-44.9739
P-value	0.0000	0.0000	0.0000	0.0000

Table 4.4: Descriptive statistics of log-returns of U.S. Treasury futures

Statistic	TU (2 Year)	FV (5 Year)	TY (10 Year)	US (30 Year)
Mean	0.0000419	0.0001083	0.0001642	0.0002401
Median	0.0000000	0.0001275	0.0002423	0.0004863
Mode	0.0000000	0.0000000	0.0000000	0.0000000
Std. Dev.	0.0009133	0.0024691	0.0038330	0.0065419
Variance	0.0000008	0.0000061	0.0000147	0.0000428
Minimum	-0.0082636	-0.0161355	-0.0201376	-0.0291463
Maximum	0.0081102	0.0169878	0.0330735	0.0466118
Kurtosis	7.9895092	3.0574585	3.1012147	2.1217655
Skewness	-0.0240245	0.0104678	0.1010453	0.0772974

Now, let us define a daily realized volatility, a measure which can be computed for a day t in a following way:

$$RVOL_t = \sqrt{RV_t}, \tag{4.4}$$

where the realized variance is defined as (McAleer & Medeiros, 2008):

$$RV_t = \sum_{i=0}^{n_t} r_{t,i}^2, \tag{4.5}$$

where $r_{t,i}$ are intraday returns collected on a day t in high-frequency. It can be shown that under the condition of no microstructure noise, which can be caused by the bid-ask spread in the financial markets, the realized variance in equation (4.5) is a consistent estimator of the integrated variance in equation (4.3) (McAleer & Medeiros, 2008). For these reasons, we measure the performance of our volatility models in forecasting the realized volatility. More specifically, we use the realized volatility calculated from the 5-minute intraday returns, which is a sampling frequency offering a decent balance between the relevance of the asymptotics and harmful effects of the microstructure noise (Andersen *et al.*, 2001).

The realized volatility series of log-returns of the studied futures are depicted in the figure 4.3. All the realized volatilities show persistent behaviour and a period of increased volatility during the global financial crisis 2007-2009. The results of the Augmented Dickey-Fuller test in table 4.5 show that we can, as in the case

of log-returns, reject the null hypothesis of a unit-root for each realized volatility series and so allow us to estimate the underlying DGP. The descriptive statistics are summarized in the table 4.6. The high excess kurtosis is again a signal of a leptokurtic distribution. The positive skewness measure, which is common for each of the studied realized volatility series, means that high volatility extremes are vastly more likely than the opposite. Mean level of volatility increases with the longer term to maturity of the underlying security, which is connected with a greater uncertainty.

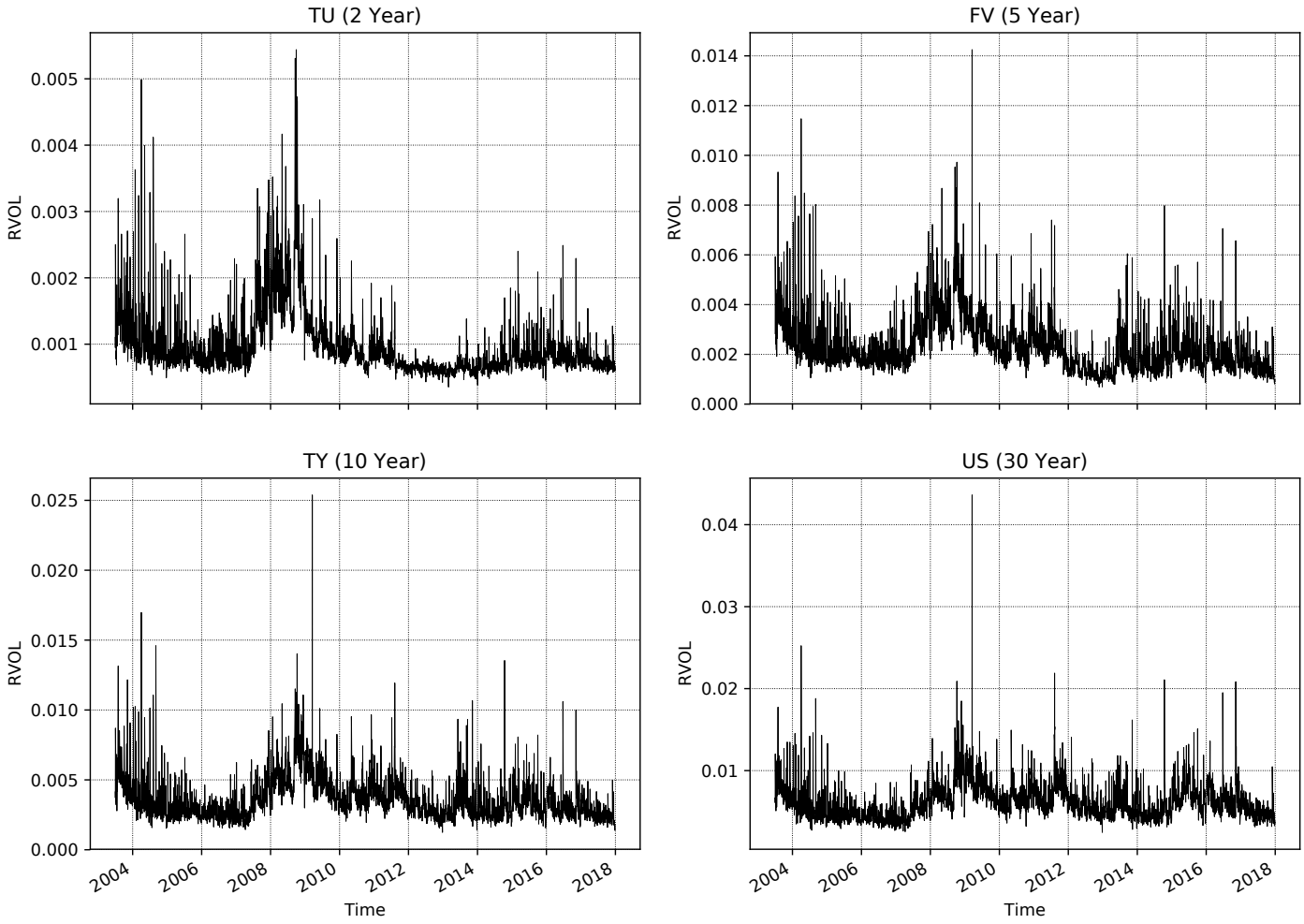

Figure 4.3: Realized volatility of log-returns of U.S. Treasury futures

4.2.2 Volatility models

In the following subsections are described the individual volatility models used for forecasting the realized volatility of U.S. Treasury futures. The aim of this study is not to achieve the most accurate forecasts of the realized volatility possible, but rather to assess whether any improvements to the forecast accuracy can be made

Table 4.5: Augmented Dickey-Fuller test results for the realized volatility of log-returns of U.S. Treasury futures

	TU (2 Year)	FV (5 Year)	TY (10 Year)	US (30 Year)
Test Statistic	-3.7401	-4.5063	-4.5752	-4.3997
P-value	0.0036	0.0002	0.0001	0.0003

Table 4.6: Descriptive statistics of realized volatility of log-returns of U.S. Treasury futures

Statistic	TU (2 Year)	FV (5 Year)	TY (10 Year)	US (30 Year)
Mean	0.0009713	0.0023081	0.0035978	0.0062063
Median	0.0008033	0.0020238	0.0031799	0.0056792
Mode	0.0003544	0.0006875	0.0012428	0.0024624
Std. Dev.	0.0005017	0.0011171	0.0015791	0.0023289
Variance	0.0000003	0.0000012	0.0000025	0.0000054
Minimum	0.0003544	0.0006875	0.0012428	0.0024624
Maximum	0.0054402	0.0142512	0.0253829	0.0436211
Kurtosis	13.7238906	10.6326379	16.1089503	23.0272817
Skewness	3.0514297	2.3825485	2.6280856	2.7244267

by combining the forecasts in a common application. Therefore, in the selection of the individual volatility models, we limit ourselves to the set of the most widely known and applied models. The models are presented approximately in order from the simplest one to the slightly more complicated ones.

Historical Volatility

The first forecast of the realized volatility in our application is a simple projection of the historical volatility h steps ahead:

$$\sigma_{T+h} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T} r_t^2}, \tag{4.6}$$

where T is the length of the training sample and r_t is the t-th day log-return. We acknowledge that forecasting realized volatility using the historical volatility on itself is a naive approach. Nevertheless, we find its inclusion into the set of individual volatility forecasts interesting in that it allows us to observe how the combining methods can deal with the presence of such model, which tends to forecast rather inaccurately. Also, it is a useful performance benchmark for the other volatility models. The h-steps-ahead forecasts using this method is constant for an arbitrary $h \geq 1$.

RiskMetrics

One of the most widely applied volatility estimators in practice is the RiskMetrics, the product of J.P. Morgen. RiskMetrics is based on the exponentially weighted moving average (EWMA) process, which supposedly reflects the finite memory of the financial markets (Pafka & Kondor, 2001). The model is following:

$$\sigma_{t+1} = \sqrt{(1-\lambda)\sum_{\tau=0}^{\infty} r_{t-\tau}^2}, \tag{4.7}$$

where the recommended value of the parameter λ is 0.94. The equation (4.7) can be further rewritten as:

$$\sigma_{t+1} = \sqrt{(1-\lambda)r_t^2 + \lambda\sigma_t^2}.$$

For initialization, we the use the historical volatility (4.6). Because $\mathbb{E}(r_t^2) = \sigma_{t+1}^2$, it turns out, yet again, that h-steps-ahead forecasts are constant for an arbitrary $h \geq 1$.

HAR

The next individual forecasting model we use is the notoriously known Heterogeneous Autoregressive model of Realized Volatility (HAR) of Corsi (2009). The motivation behind this model is that there are investors with different time horizons and hence volatility can be decomposed into multiple (three) components corresponding to different periods (daily, weekly and monthly), which all in its part influence the future volatility expectations. The HAR model is defined as follows:

$$RVOL_{t+1}^{(d)} = \beta_0 + \beta_d RVOL_t^{(d)} + \beta_w RVOL_t^{(w)} + \beta_m RVOL_t^{(m)} + \epsilon_{t+1}, \tag{4.8}$$

where the individual components are derived from the past realized volatilities as:

$$RVOL_t^{(i)} = \frac{1}{i}\sum_{\tau=1}^{i} RVOL_{t+1-\tau}.$$

The parameters d, w, m are assumed to be equal to 1, 5, 22 respectively. Finally, ϵ_{t+1} is the serially independent error term. The HAR model can be easily estimated via OLS. The h-steps-ahead forecast of the realized volatility can be obtained recursively.

GARCH

A traditional benchmark model in econometric volatility forecasting literature is the famous GARCH(1,1). First, Engle (1982) introduced a class of Autoregressive Conditional Heteroskedasticity (ARCH) processes, which distinguish between the constant unconditional variance and a conditional variance, which is allowed to change

in time. This class was later extended into Generalized Autoregressive Conditional Heteroskedasticity (GARCH) processes, which allow for a greater flexibility and a longer memory in volatility (Bollerslev, 1986). In the simplest form, the GARCH(1,1) is defined as follows:

$$\epsilon_t|\psi_{t-1} \sim \mathcal{N}(0, \sigma), \tag{4.9}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \beta_1\sigma_{t-1}^2, \tag{4.10}$$

where

$$\alpha_0 > 0, \quad \alpha_1 \geq 0, \quad \beta_1 \geq 0.$$

Here the conditional variance σ_t^2 is the function of past sample variances ϵ_{t-1}^2 and lagged conditional variances σ_{t-1}^2 as well. In order for the wide sense stationarity property to hold, it is sufficient that $\alpha_1 + \beta_1 < 1$. The parameters of the process α_0, α_1 and β_1 can be estimated using MLE (Bollerslev, 1986). Again, the h-steps-ahead volatility forecast can be obtained recursively.

VAR

The final individual model used in this thesis is the Vector Autoregressive (VAR) model. Its use for the multivariate modelling and forecasting of realized volatility is inspired by Andersen *et al.* (2003), who use a fractionally-integrated VAR to forecast logarithmic realized volatility of spot exchange rate. We hypothesise, that volatility of U.S. Treasury futures may be interrelated, and that broadening the information sets for more than just a single series could improve the forecasting performance. As in Andersen *et al.* (2003), we use 5 lags in each equation, to allow the VAR to capture even complicated structures in the realized volatility. The model is following:

$$\boldsymbol{y}_t = \boldsymbol{A}_0 + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \boldsymbol{A}_3\boldsymbol{y}_{t-3} + \boldsymbol{A}_4\boldsymbol{y}_{t-4} + \boldsymbol{A}_5\boldsymbol{y}_{t-5} + \boldsymbol{\epsilon}_t, \tag{4.11}$$

where $\boldsymbol{A}_0$ is the $(r \times 1)$ vector of intercepts, $\boldsymbol{A}_1$, …, $\boldsymbol{A}_5$ are the $(r \times r)$ matrices of regression coefficients, $\boldsymbol{y}_t$ is the $(r \times 1)$ vector containing realized volatilities of r different futures at time t and $\boldsymbol{\epsilon}_t$ is the $(r \times 1)$ vector of independently, normally distributed residuals. In building the model, we got inspired by the thought of Bates & Granger (1969), that a combination of forecasts based on different model specifications or information sets can be superior in performance to all the combined forecasts individually. Therefore, we have decided to include all the VARs with different combinations of available futures (TU, FV, TY, US) among the individual volatility forecasting models. In total, we work with $\sum_{r=1}^{4}\binom{4}{r} = 15$ different VAR model specifications, from which $\sum_{r=0}^{3}\binom{3}{r} = 8$ can be used to forecast the realized volatility of each different future. Since all the considered VARs are in the reduced

form, they can be estimated via OLS equation by equation. The usual diagnostic checks were applied to the VARs estimated on the whole sample. All the roots of the characteristic polynomials lie outside the unit circle, implying that the VARs are stable. Also, the sample autocorrelation functions of residuals of respective VARs do not reveal any leftover information. The h-steps-ahead forecasts of realized volatilities can be obtained recursively.

4.2.3 Individual Forecasts

The volatility models described above were estimated on the U.S. Treasury futures realized volatility using a rolling window of length 1000 and the 1, 5 and 22-steps-ahead forecasts were obtained leaving us with 2635, 2631 and 2614 forecast observations respectively to be combined for each Treasury. The individual 1, 5 and 22-steps-ahead volatility forecasts are depicted in the figures 4.4, 4.5 and 4.6 respectively. The forecast performance is summarized in the table 4.7, where the measures used are defined further bellow in the section 5.1.

From both the table and figures it is apparent that the Historical Volatility forecasts are substantially worse than the forecasts of other models as it was expected. The RiskMetrics and GARCH model forecasts show similar performance. They both underpredict the realized volatility the of TU futures in the period 2011–2015. The most accurate individual forecasting model overall in our application is the HAR model, followed by the VAR models. Regarding the different VAR model specifications, it shows that simple specifications such the one variable VAR, which basically corresponds to an AR(5) model, is more accurate in forecasting the next day realized volatility (h=1), while the complex specifications such the four variable VAR are more accurate in forecasting the realized volatility in longer horizons (h=22). The finding supports the claim that past realized volatilities of U.S. Treasury futures also carry some information about the long-term future realized volatility of U.S. Treasury futures based on bonds with different maturities. For other futures than the TU (2 Year), all the individual forecasts, excluding the Historical Volatility, are very much comparable. Across all the models and futures, the accuracy of forecasts naturally tends to decrease as the forecast horizon (h) increases.

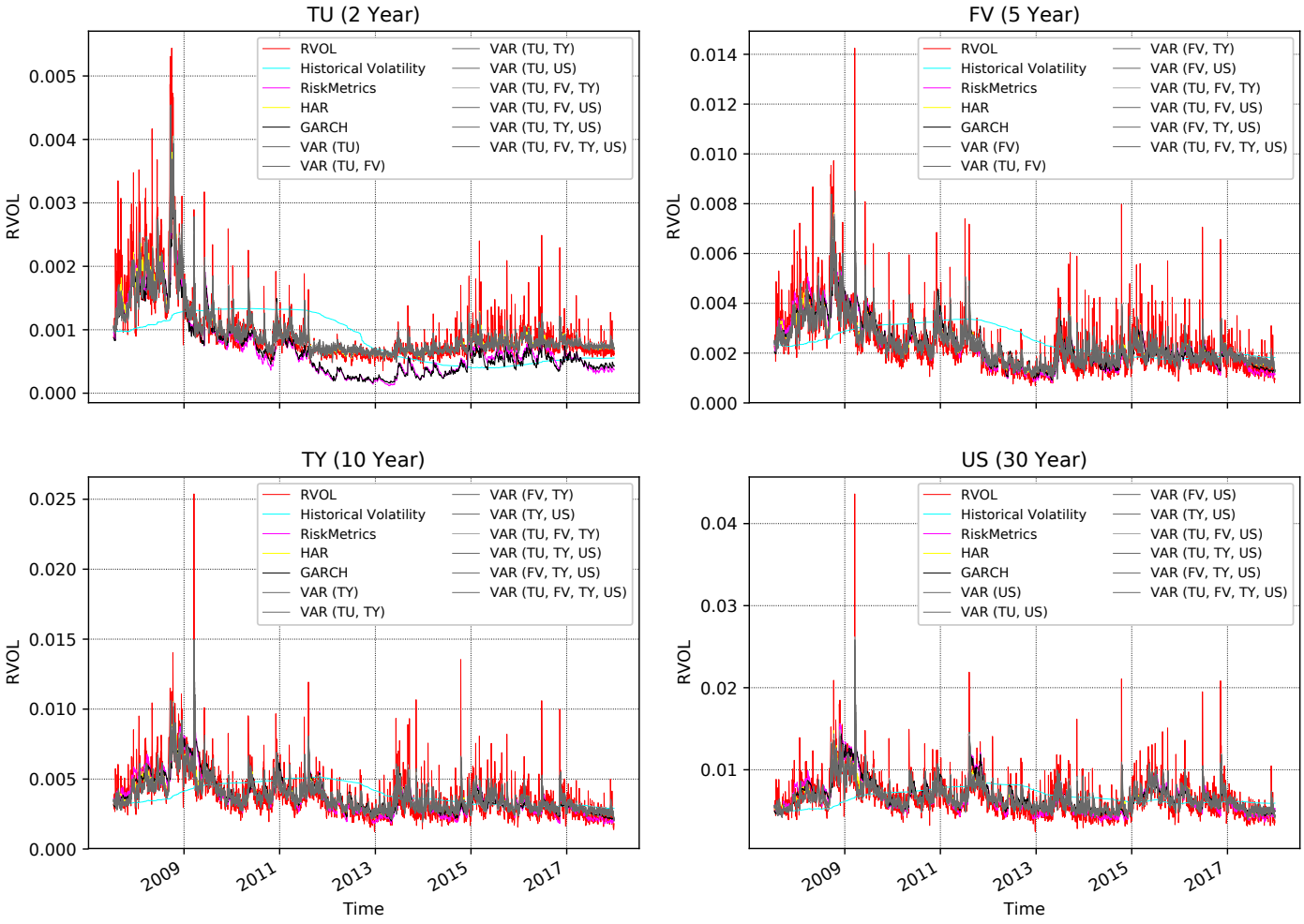

Figure 4.4: Individual 1-step-ahead forecasts of the realized volatility

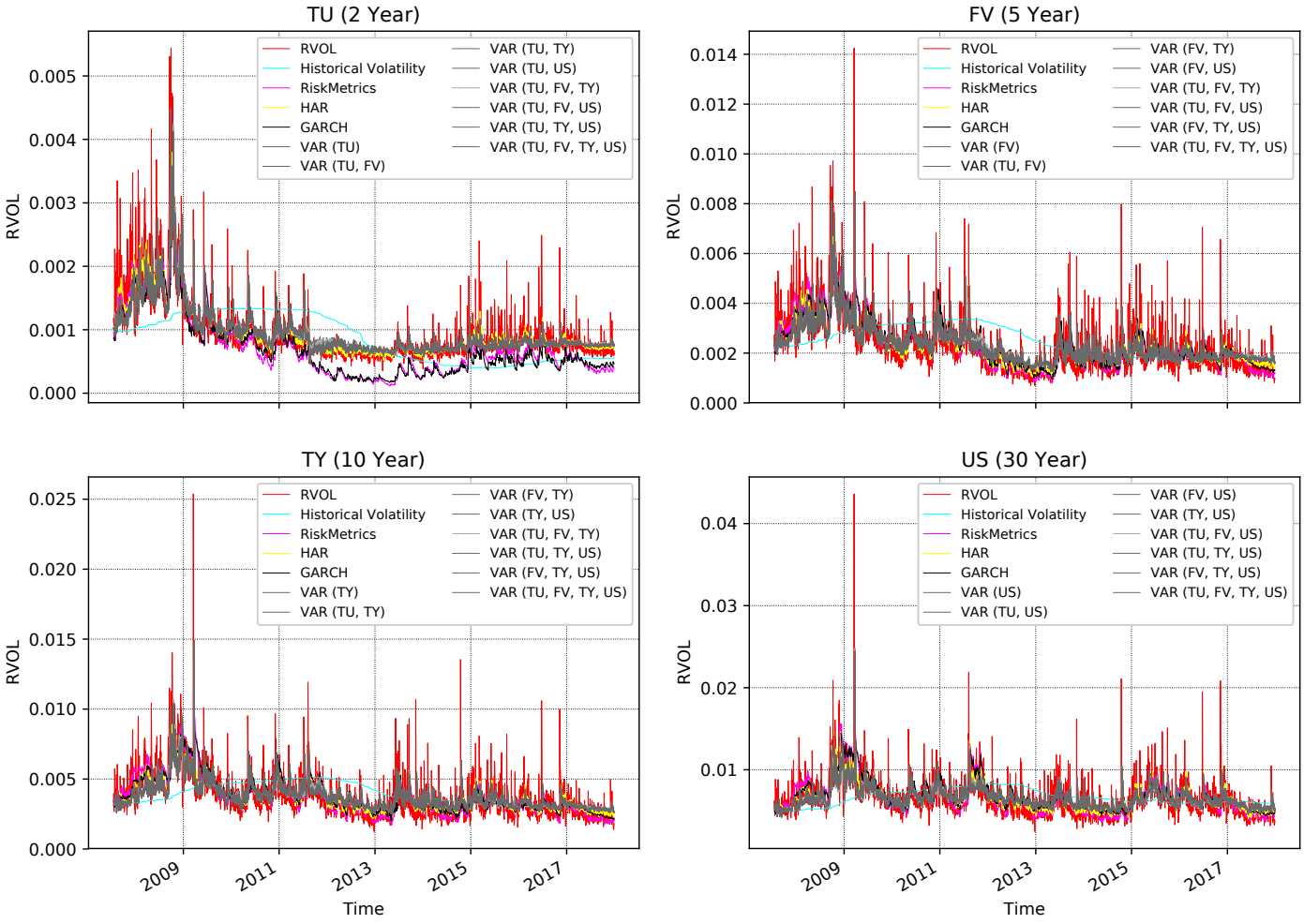

Figure 4.5: Individual 5-steps-ahead forecasts of the realized volatility

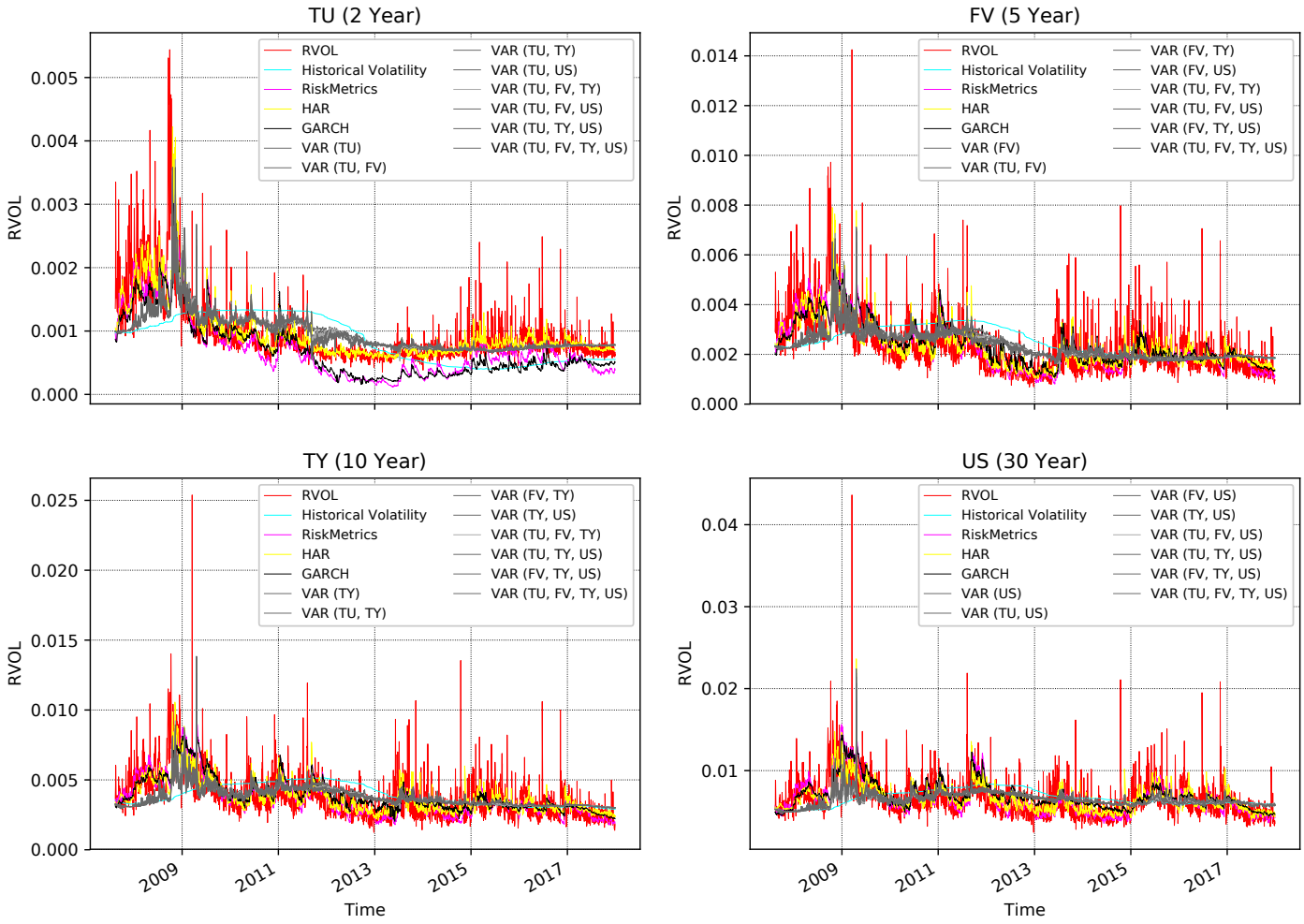

Figure 4.6: Individual 22-steps-ahead forecasts of the realized volatility

Table 4.7: Forecast performance (measured in terms of RMSE, MAE and MAPE) of indindividual volatility models in h-steps-ahead forecasting of the realized volatility of U.S. Treasury futures log-returns

Future	Volatility Model	h = 1			h = 5			h = 22		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
TU (2 Year)	Historical Volatility	5.50	4.02	40.95	5.53	4.04	41.18	5.59	4.09	41.95
	RiskMetrics	3.96	2.96	33.11	4.13	3.07	33.84	4.50	3.23	34.49
	HAR	2.72	1.49	13.44	3.13	1.69	15.15	3.86	2.04	18.05
	GARCH	4.08	3.03	33.50	4.24	3.14	34.19	4.64	3.37	35.95
	VAR (TU)	2.76	1.51	13.67	3.32	1.86	17.06	4.52	2.78	27.11
	VAR (TU, FV)	2.79	1.53	13.95	3.34	1.88	17.44	4.45	2.72	26.53
	VAR (TU, TY)	2.80	1.54	13.96	3.38	1.90	17.45	4.52	2.78	26.94
	VAR (TU, US)	2.80	1.52	13.75	3.34	1.85	16.90	4.42	2.67	25.66
	VAR (TU, FV, TY)	2.80	1.56	14.28	3.37	1.92	17.82	4.51	2.79	27.17
	VAR (TU, FV, US)	2.81	1.54	14.03	3.35	1.87	17.13	4.42	2.67	25.78
	VAR (TU, TY, US)	2.80	1.55	14.04	3.35	1.87	17.13	4.43	2.69	25.76
	VAR (TU, FV, TY, US)	2.80	1.56	14.30	3.34	1.89	17.53	4.42	2.70	26.16
FV (5 Year)	Historical Volatility	11.88	8.85	44.43	11.93	8.89	44.66	12.08	9.00	45.48
	RiskMetrics	8.19	5.17	21.58	8.65	5.60	23.53	9.55	6.36	27.13
	HAR	7.53	4.74	20.51	8.34	5.29	22.72	9.57	6.26	27.23
	GARCH	8.22	5.29	22.81	8.65	5.72	24.98	9.53	6.62	30.47
	VAR (FV)	7.65	4.84	21.30	8.70	5.75	26.06	10.62	7.50	36.24
	VAR (TU, FV)	7.70	4.90	21.65	8.70	5.75	26.26	10.41	7.23	35.08
	VAR (FV, TY)	7.73	4.93	21.76	8.76	5.85	26.57	10.57	7.53	36.42
	VAR (FV, US)	7.72	4.86	21.32	8.70	5.71	25.70	10.43	7.29	34.84
	VAR (TU, FV, TY)	7.72	4.98	22.20	8.70	5.84	26.98	10.37	7.27	35.54
	VAR (TU, FV, US)	7.73	4.91	21.68	8.70	5.74	26.15	10.36	7.22	34.88
	VAR (FV, TY, US)	7.73	4.97	22.05	8.66	5.79	26.56	10.31	7.28	35.14
	VAR (TU, FV, TY, US)	7.73	5.00	22.31	8.64	5.79	26.84	10.23	7.17	35.04
TY (10 Year)	Historical Volatility	16.47	12.05	34.30	16.55	12.11	34.47	16.79	12.31	35.13
	RiskMetrics	12.00	7.58	19.36	12.78	8.23	21.06	14.10	9.37	24.20
	HAR	10.96	6.93	18.46	12.12	7.78	20.50	13.73	9.10	24.23
	GARCH	11.93	7.66	20.15	12.56	8.26	21.90	13.76	9.51	26.16
	VAR (TY)	11.16	7.10	19.00	12.61	8.36	22.78	15.05	10.66	30.07
	VAR (TU, TY)	11.31	7.19	19.20	12.69	8.37	22.81	14.94	10.42	29.36
	VAR (FV, TY)	11.25	7.17	19.23	12.66	8.41	22.96	14.91	10.54	29.83
	VAR (TY, US)	11.20	7.12	19.01	12.57	8.28	22.44	14.83	10.37	28.99
	VAR (TU, FV, TY)	11.29	7.26	19.54	12.64	8.46	23.26	14.81	10.37	29.44
	VAR (TU, TY, US)	11.34	7.23	19.33	12.67	8.38	22.80	14.84	10.34	29.09
	VAR (FV, TY, US)	11.27	7.22	19.44	12.60	8.41	23.11	14.72	10.34	29.23
	VAR (TU, FV, TY, US)	11.31	7.30	19.67	12.59	8.44	23.30	14.71	10.30	29.25
US (30 Year)	Historical Volatility	25.10	18.32	29.13	25.22	18.41	29.27	25.60	18.73	29.78
	RiskMetrics	18.85	12.16	17.81	20.41	13.29	19.39	22.48	15.06	22.06
	HAR	17.39	10.93	16.39	19.38	12.28	18.24	21.74	14.49	21.69
	GARCH	18.75	12.37	18.76	20.06	13.40	20.37	21.96	15.46	24.13
	VAR (US)	17.77	11.17	16.78	20.08	13.10	19.89	23.45	16.55	25.69
	VAR (TU, US)	17.94	11.22	16.81	20.16	13.05	19.71	23.52	16.42	25.29
	VAR (FV, US)	17.85	11.19	16.81	20.09	13.03	19.69	23.38	16.32	25.11
	VAR (TY, US)	17.77	11.17	16.80	19.96	13.01	19.67	23.00	16.11	24.82
	VAR (TU, FV, US)	18.02	11.37	17.14	20.26	13.24	20.19	23.50	16.43	25.44
	VAR (TU, TY, US)	18.02	11.36	17.16	20.07	13.22	20.21	22.88	16.07	25.10
	VAR (FV, TY, US)	17.84	11.32	17.14	19.83	13.13	20.16	22.52	15.85	24.83
	VAR (TU, FV, TY, US)	17.99	11.45	17.40	19.96	13.27	20.48	22.71	15.98	25.11

Note: The RMSE and MAE measures are scaled up by the order of 10^4.

Chapter 5

Forecast Performance Assessment

This main purpose of this chapter is to present and asses the results from both of our empirical applications. First, we define the individual measures of the forecast accuracy we use to asses the performance of the forecast combinations. Secondly, we describe the DM test, which is used for testing our hypotheses. Subsequently all the results from both of the applications are presented. Finally, we rank the forecast combination methods according to their overall forecast accuracy.

5.1 Measures of the Forecast Accuracy

The most common way of comparing accuracy in forecasting literature is via different accuracy measures. There exists plenty of measures in the literature, each with its own pros and cons and allowing us to view the forecast performance of a given model or method on a given dataset from a different perspective. Here we introduce the three of the most common measures, which are applied in our empirical application. For a broader review of possible measures of forecasts accuracy and their properties see e.g. Hyndman & Koehler (2006).

5.1.1 RMSE

The Root Mean Square Error (RMSE) is the most popular and frequently applied forecast accuracy measure between both practitioners and academicians (Armstrong & Collopy, 1992). It is defined as:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - f_t)^2}, \tag{5.1}$$

where y_t is the true outcome, f_t is its forecast and T is the length of the (out-of-sample) time series. It is usually preferred over the Mean Square Error (MSE) as it

is on the same scale as the data from which it is computed (Hyndman & Koehler, 2006). (Armstrong & Collopy, 1992) argue that RMSE is relevant to decision making but not a reliable measure of forecasts accuracy.

5.1.2 MAE

The Mean Absolute Error (MAE) is defined as:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|y_t - f_t|, \tag{5.2}$$

where y_t is the true outcome, f_t is its forecast and T is the length of the (out-of-sample) time series. The main argument for its use over the RMSE (5.3) is that it is generally less sensitive to outliers (Hyndman & Koehler, 2006).

5.1.3 MAPE

The Mean Absolute Percentage Error (MAPE) is defined as:

$$MAPE = \frac{100}{T^*}\sum_{t\in\{1,\dots,T|y_t\neq 0\}}\left|\frac{y_t - f_t}{y_t}\right|, \tag{5.3}$$

where y_t is the true outcome, f_t is its forecast and T^* is the length of the (out-of-sample) time series reduced by the number of observations where the outcome equals 0. The main advantage of MAE is that it is, unlike the two previously described measures, a unit-free measure, which allows for cross-series comparisons. The disadvantage is that it is not suitable for series where a large share of observations is equal to 0 (Hyndman & Koehler, 2006). Another disadvantage is that this measure favours models which tend to underpredict versus those which tend to overpredict, i.e. it is an asymmetric measure of forecast accuracy (Armstrong & Collopy, 1992).

5.2 DM Test

Diebold-Mariano test of equal forecast accuracy (DM test), was introduced by Diebold & Mariano (2002). It is used to test the null hypothesis of equal expected loss (equal forecast accuracy) between a given pair of forecasts. The original intention behind the test was to developed a tool for statistical forecast comparisons in model-free environments (Diebold, 2015). This is exactly our case, as a considerable share of forecast combinations in our study is not based on a proper statistical model. We therefore see the DM test as appropriate tool in our situation and we use it to test our hypotheses.

Let us denote $d_{1,2;t}$ a loss differential defined as:

$$d_{1,2;t} = L(e_{1,t}) - L(e_{2,t})$$

where, the e_1, e_2 are forecast errors of the same given quantity at time t and $L(.)$ is a loss functions. For the testing in our study, we use the quadratic loss:

$$L(e_t) = e_t^2.$$

Assuming that the following conditions, corresponding to the covariance stationarity of the loss differentials, hold, is sufficient for the validity of the DM test:

- $\mathbb{E}(d_{1,2;t}) = \mu, \quad \text{for } \forall t$
- $Var(d_{1,2;t}) = \sigma^2 < \infty, \quad \text{for } \forall t$
- $Cov(d_{1,2;t}, d_{1,2;t-\tau}) = \gamma(\tau), \quad \text{for } \forall t, \tau$

Assuming the conditions hold, under the null hypothesis of equal expected loss (equal forecast accuracy):

$$H_0 : \mathbb{E}(d_{1,2;t}) = 0, \quad \text{for } \forall t,$$

it holds that:

$$DM_{1,2} = \frac{\sum_{t=1}^{T} d_{1,2;t}}{T\hat{\sigma}_{1,2}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $DM_{1,2}$ denotes the test statistic and $\hat{\sigma}_{1,2}$ is a consistent estimator of the loss differential standard error (Diebold, 2015). In order to deal with the autocorrelation of the loss differentials, induced by e.g. the misspecification of the forecast models, one can use the heterogeneity and autocorrelation consistent (HAC) standard errors of Newey & West (1986), with appropriate amount of lags. The use of 4 lags is a natural choice in our case, as we work the quarterly data series from the ECB SPF application. The DM statistic can be easily obtained by regressing the loss differential on an intercept.

In reality, no pair of forecast errors is likely ever to give truly covariance stationary loss differentials. Nevertheless, we rely on an approximate validity of the covariance stationarity condition. As Diebold (2015) notes, the forecasts are often based on similar information sets and may share the non-stationary components in the errors. Therefore, the non-stationarity may cancel out and not translate into the loss differentials.

5.3 ECB Survey of Professional Forecasters

The forecast combinations methods described in the chapter 3 were estimated on the the forecasts of the real GDP growth, harmonised inflation and unemployment rate with 1 and 2 year horizons from the ECB SPF. The estimation or training of the combination methods was done on a rolling basis, using windows of lengths 25, 35 and 45 observations and the 1-step-ahead out-of-sample forecast combinations were obtained.

The figures 5.1, 5.2 and 5.3 show the forecasts of several combinations methods and the target variable for the respective lengths of the training window: 25, 35, 45 observations (quarters). For clarity, only the best performing forecast combinations on a given dataset from each class, as divided into sections in the methodology chapter 3, were selected to be displayed in the figures. By the best performing forecast combination, we understand the combination that a on given dataset has the lowest average rank from the separate rankings of the combination methods according to the accuracy measures RMSE, MAE and MAPE. Across the window lengths, it appears that the most erratic forecasts are from the shrinkage methods, mostly the Kappa-Shrinkage. The reason being that it gives a weighted average of equal weights and an OLS forecast of weights, which we later show give very inaccurate forecasts in the ECB SPF application overall. Also, even the best of the Alternative (machine learning) forecast combinations show a very erratic behaviour, especially in the case of the harmonised inflation and the unemployment rate. Naturally, it is because these methods require bigger samples for training than the ECB SPF application can provide. The plot of the 2 year horizon forecasts of the real GDP growth for the length of the window 25 shows that all of the combining methods completely failed in the 2 year horizon forecasting of the global financial crisis slump. However, since none of the individual forecasts did anticipate the crisis (see figure 4.2), we can infer that the information was simple not in the data for the forecast combinations to figure out.

The tables 5.1, 5.2 and 5.3 summarize the forecast performance of the forecast combination methods in the described accuracy measures for the sample rolling window lengths of 25, 35 and 45 observations respectively. Firstly. note the main reason for the dramatic improvement of measures in the window length 45 vs. 35 is explained by the fact that the testing sample for the window of length 45 does no longer contain most of the financial crisis (see figure 5.3. As for the individual forecasts, the accuracy measures are generally lower for the 1 year horizons in comparison to the 2 year horizon. The class of simple forecast combinations, including the equal weights, shows very competitive results across all the ECB SPF datasets. The only exception being the group of Granger-Ramanathan OLS based forecast combinations, which on

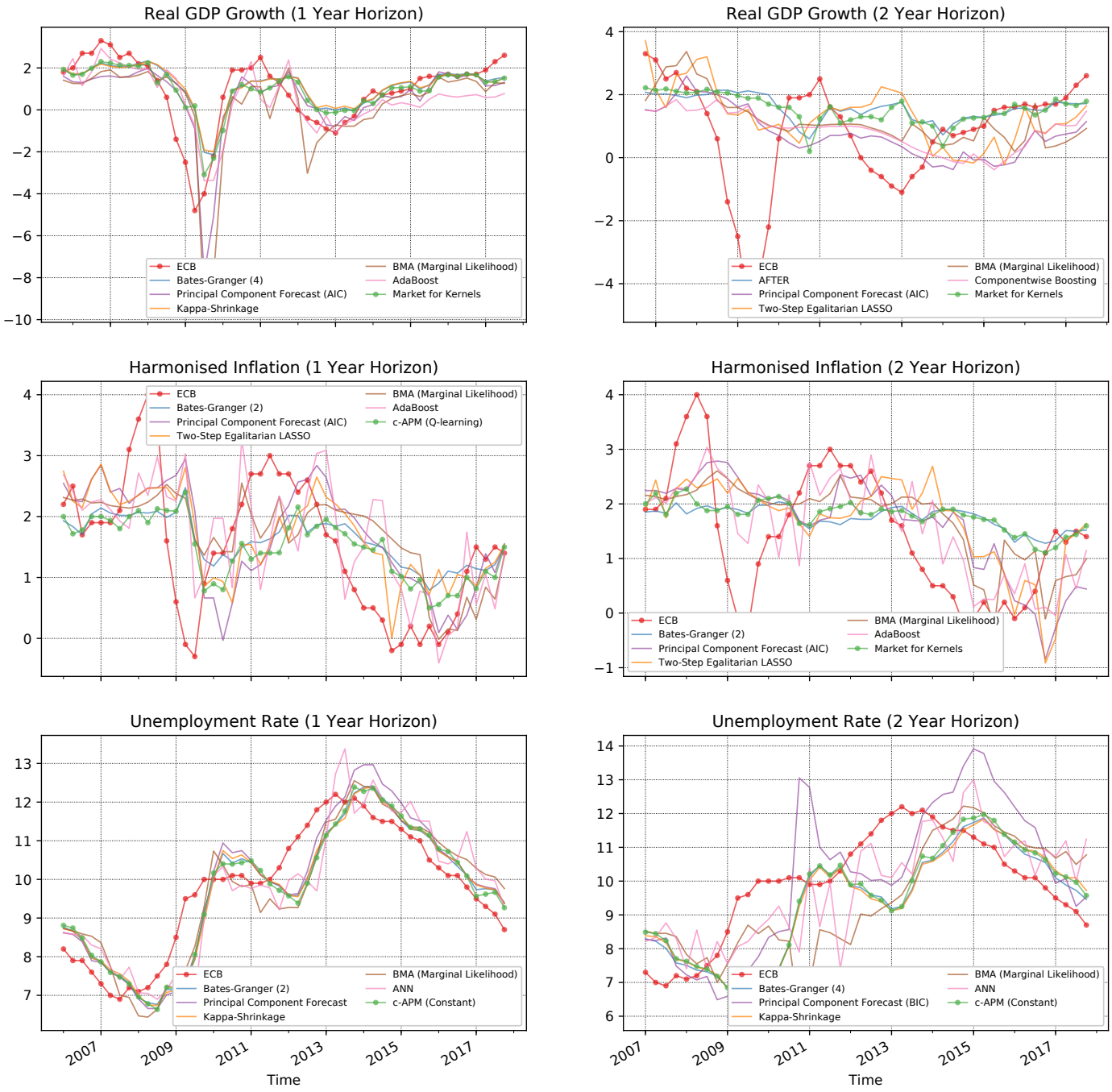

Figure 5.1: Best combinations of forecasts from the ECB SPF, trained on a rolling window of length: 25

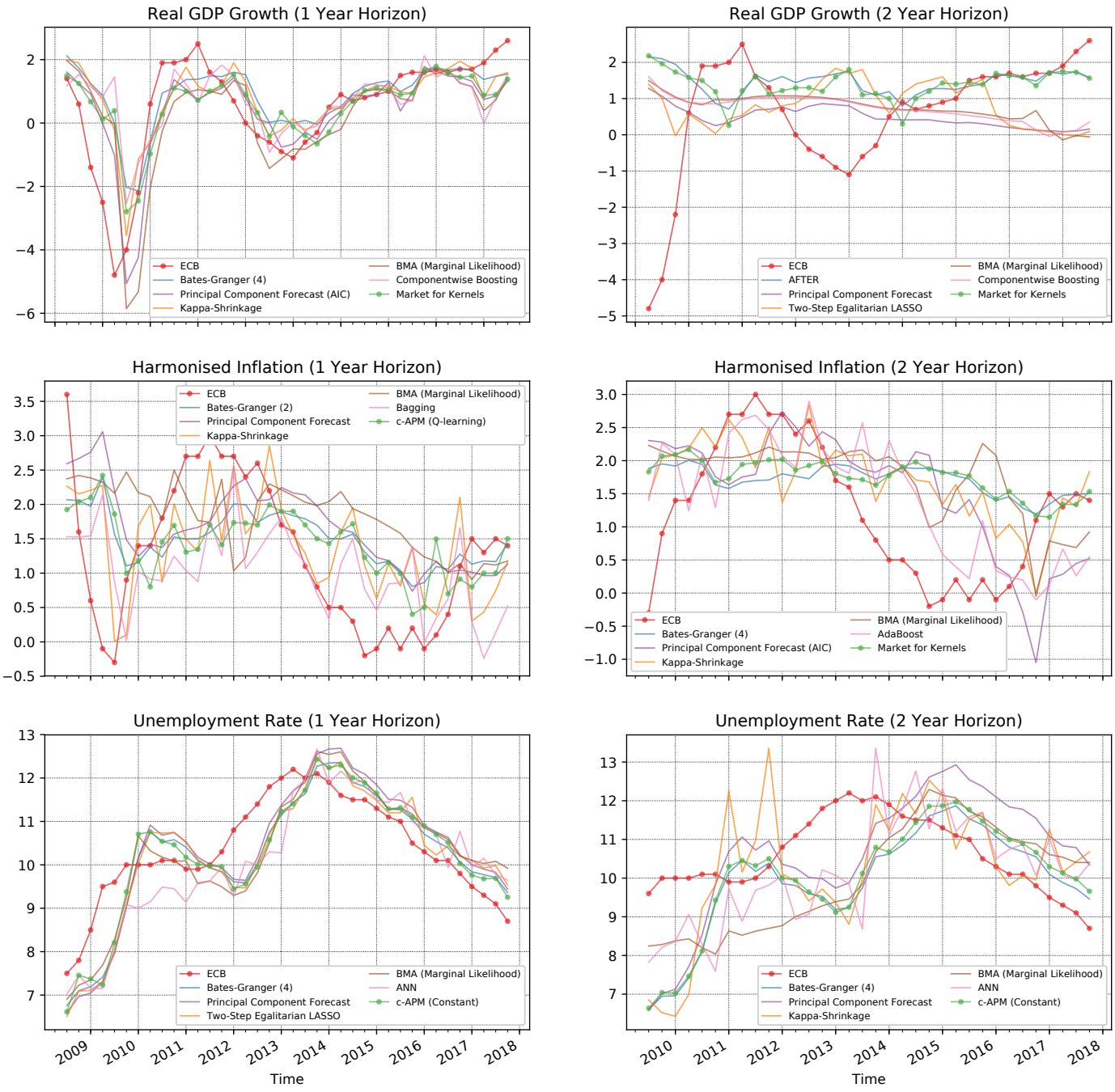

Figure 5.2: Best combinations of forecasts from the ECB SPF, trained on a rolling window of length: 35

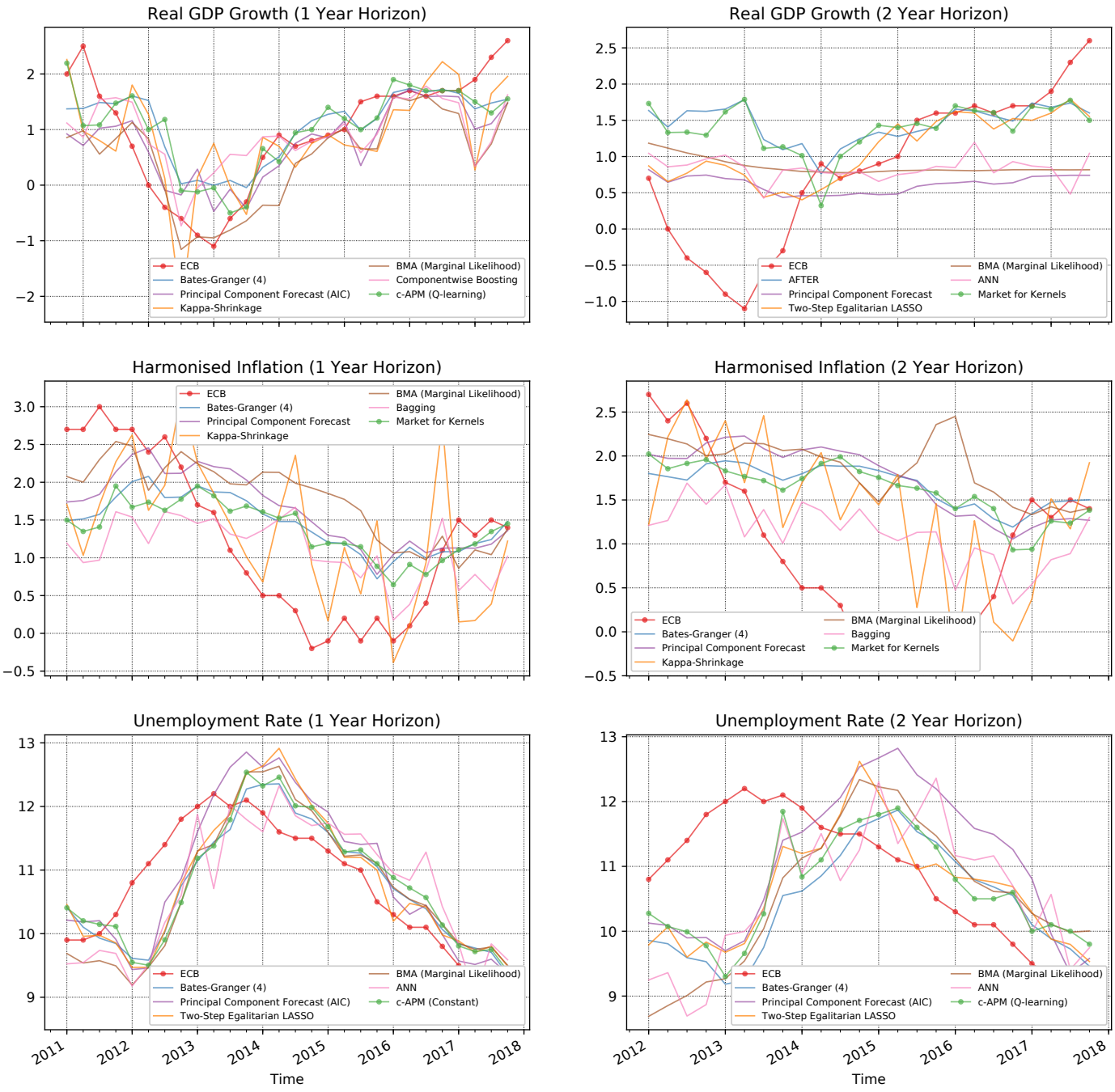

Figure 5.3: Best combinations of forecasts from the ECB SPF, trained on a rolling window of length: 45

the other hand, perform among the worse methods. The reason is that the training sample is too short and wide (about 20 individual forecasts) for the estimation. The shortness of the training sample also reflect in the worst overall performances of the EP-NN and the Bagging method. The whole class of factor analytic combinations as well as the artificial prediction markets show a solid performance in relative to the remaining forecast combinations. Regarding the comparison with the individual forecasts, the forecast combinations only rarely beat in the selected measures the best of the individuals (e.g. principal components forecasts of the real GDP growth in 2 year horizon). However, most of the methods from the simple, factor analytic and artificial prediction market class achieve lower accuracy measures than the median individual forecaster.

The first hypothesis we test in this study is that the equal weights (simple average) combination method forecasts of the macroeconomic variables from the ECB SPF are equally accurate as the forecasts of the other described forecast combinations. The table 5.4 shows the p-values from the DM test. We can reject the null hypthesis of equal forecast accuracy in most of the datasets for the Granger-Ramanathan, majority of the alternative methods, BMA based on the predictive likelihood and for the shrinkage methods on some of the datasets, as it turns out that these methods forecast significantly worse than the equal weights. On the other hand, we can reject the hypothesis for the Bates-Granger methods, namely the Bates-Granger (4), in most of the cases and for the APM methods in several cases, as it shows these forecasts are significantly better than the forecasts based on equal weights. The forecast loss differentials between the equal weights and the factor analytic and some of the other simple forecast combinations such as the AFTER, Median Forecast and PEW are generally not significantly different from zero.

Our second hypothesis is that the newly proposed method, the Market for Kernels, forecasts of the macroeconomic variables from the ECB SPF are equally accurate as the forecasts of the other described forecast combinations. The table 5.5 shows the p-values from testing the hypothesis using the DM test. In the vast majority of cases, we can either reject the hypothesis in favour of the Market for Kernels or we do not have enough evidence to claim the Market for Kernels is significantly better or worse than other methods. There are only rare cases where the Market for Kernel is found to perform significantly worse than some other method. For example the Two-step Egalitarian LASSO and the Q-learning c-APM are shown to significantly outperform the Market for Kernels in combining the forecasts of the rate of unemployment in a 2 year horizon, using the training window of length 45.

Table 5.1: Performance of forecast combinations of ECB SPF forecasts using the training window of the length: 25

Class	Forecast Combination Method	RGDP						HICP						UNEM					
		1Y			2Y			1Y			2Y			1Y			2Y		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Simple	Equal Weights	1.22	0.82	58.77	2.06	1.33	109.71	1.01	0.81	246.30	1.19	0.99	331.04	0.76	0.63	6.57	1.54	1.23	12.31
	Bates-Granger (1)	1.21	0.81	57.97	2.06	1.32	109.35	1.00	0.81	243.85	1.18	0.98	331.12	0.76	0.63	6.57	1.53	1.23	12.33
	Bates-Granger (2)	1.18	0.79	55.95	2.04	1.30	107.60	0.97	0.78	236.13	1.16	0.97	320.78	0.75	0.60	6.32	1.52	1.21	12.15
	Bates-Granger (3)	1.21	0.81	58.44	2.06	1.32	109.56	1.01	0.81	245.32	1.19	0.98	331.07	0.76	0.63	6.57	1.53	1.23	12.32
	Bates-Granger (4)	1.18	0.79	55.86	2.05	1.31	107.88	0.99	0.80	239.58	1.16	0.96	325.61	0.74	0.61	6.33	1.48	1.17	11.61
	Bates-Granger (5)	1.18	0.79	56.99	2.06	1.31	108.56	0.97	0.79	234.03	1.17	0.97	324.62	0.75	0.61	6.37	1.52	1.23	12.34
	Granger-Ramanathan (1)	3.08	1.85	199.19	4.00	2.73	231.32	1.73	1.26	230.41	2.37	1.79	365.16	1.53	1.06	10.67	3.21	2.28	22.63
	Granger-Ramanathan (2)	2.44	1.61	150.00	4.29	2.99	247.16	1.60	1.18	251.36	1.94	1.59	327.65	1.32	0.94	9.35	3.23	2.41	23.81
	Granger-Ramanathan (3)	3.74	2.19	223.03	4.69	2.94	280.74	2.40	1.67	416.03	3.03	2.10	390.49	1.67	1.16	11.62	3.26	2.31	22.48
	AFTER	1.19	0.80	54.91	2.03	1.29	104.69	1.00	0.81	238.68	1.18	0.97	330.06	0.76	0.63	6.60	1.50	1.22	12.23
	Median Forecast	1.21	0.80	57.84	2.06	1.33	110.14	1.01	0.82	245.57	1.19	0.99	333.93	0.77	0.63	6.58	1.53	1.24	12.43
	Trimmed Mean Forecast	1.22	0.82	58.48	2.06	1.33	109.71	1.01	0.81	246.30	1.18	0.98	331.27	0.76	0.63	6.57	1.54	1.23	12.31
	PEW	1.55	1.05	69.75	1.90	1.35	101.88	1.08	0.87	252.26	1.19	0.90	365.18	0.80	0.67	7.13	1.47	1.32	13.64
Factor An.	Principal Component Forecast	1.29	0.91	60.86	1.90	1.50	97.90	1.03	0.84	249.45	1.12	0.97	298.20	0.79	0.68	6.99	1.57	1.37	13.77
	Principal Component Forecast (AIC)	1.28	0.89	53.76	1.91	1.45	97.83	1.08	0.90	218.37	1.12	0.92	237.67	0.87	0.73	7.42	1.64	1.41	14.11
	Principal Component Forecast (BIC)	1.32	0.90	53.63	1.93	1.48	99.10	1.06	0.89	235.62	1.13	0.94	240.07	0.86	0.72	7.36	1.58	1.36	13.62
Shrinkage	Empirical Bayes Estimator	3.32	1.88	177.55	4.19	2.51	242.65	2.23	1.57	377.46	2.72	1.84	361.76	1.47	1.03	10.34	2.81	2.02	19.66
	Kappa-Shrinkage	1.22	0.82	58.77	2.06	1.33	109.71	1.01	0.81	246.30	1.19	0.99	331.04	0.76	0.63	6.57	1.54	1.23	12.31
	Two-Step Egalitarian LASSO	1.43	0.94	77.88	1.94	1.45	120.43	0.99	0.83	225.47	1.17	0.96	245.64	0.84	0.71	7.20	1.73	1.49	14.70
BMA	BMA (Marginal Likelihood)	1.76	1.21	89.39	1.82	1.31	92.63	1.01	0.81	217.00	1.04	0.86	232.83	0.83	0.72	7.60	1.64	1.42	14.34
	BMA (Predictive Likelihood)	2.09	1.54	160.44	2.52	1.97	196.50	1.24	1.06	294.25	1.22	0.95	303.64	2.20	1.44	14.08	3.80	2.52	23.67
Alternative	ANN	1.70	1.10	75.18	2.33	1.68	131.61	1.01	0.80	189.29	1.26	1.00	305.21	0.89	0.71	7.38	1.28	1.08	11.20
	EP-NN	4.92	3.06	312.50	14.95	8.99	924.46	6.34	4.36	1154.67	132.67	24.02	1626.42	2.12	1.45	14.92	9.10	5.45	52.31
	Bagging	1.95	1.69	99.12	1.97	1.69	106.60	1.79	1.47	95.46	1.73	1.32	119.56	7.60	7.50	76.97	9.59	9.48	95.46
	Componentwise Boosting	1.40	0.97	65.33	1.84	1.38	95.48	1.01	0.82	230.84	1.04	0.86	239.04	1.49	1.34	13.74	1.60	1.42	14.33
	AdaBoost	1.28	0.97	70.47	2.15	1.69	116.79	1.00	0.79	191.68	0.88	0.70	160.66	1.35	1.17	11.84	1.72	1.46	14.34
APM	c-APM (Constant)	1.22	0.81	55.17	2.05	1.30	107.53	0.98	0.80	232.11	1.17	0.98	331.59	0.77	0.62	6.49	1.50	1.21	12.17
	c-APM (Q-learning)	1.23	0.84	60.50	2.01	1.25	104.78	0.97	0.77	204.71	1.17	0.98	336.76	0.79	0.62	6.49	1.58	1.29	12.91
	Market for Kernels	1.12	0.74	51.90	1.95	1.22	98.09	0.99	0.79	215.97	1.11	0.93	317.64	0.76	0.65	6.72	1.53	1.28	12.90
	Best Individual	1.07	0.76	50.78	1.94	1.28	98.45	0.95	0.74	206.99	1.12	0.94	286.37	0.74	0.55	5.71	1.45	1.07	10.66
	Median Individual	1.29	0.88	63.45	2.07	1.34	111.20	1.04	0.83	247.12	1.20	1.01	333.22	0.80	0.66	6.82	1.58	1.29	12.87
	Worst Individual	1.49	1.00	88.23	2.28	1.54	136.86	1.19	0.97	307.70	1.37	1.17	397.37	0.96	0.77	7.97	1.84	1.52	15.06

Table 5.2: Performance of forecast combinations of ECB SPF forecasts using the training window of the length: 35

Class	Forecast Combination Method	RGDP						HICP						UNEM					
		1Y			2Y			1Y			2Y			1Y			2Y		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Simple	Equal Weights	1.32	0.88	69.06	2.07	1.28	115.91	1.03	0.85	305.50	1.14	0.96	355.80	0.81	0.67	6.58	1.60	1.29	12.15
	Bates-Granger (1)	1.31	0.88	68.19	2.07	1.28	115.55	1.03	0.85	304.27	1.14	0.96	355.67	0.81	0.67	6.59	1.60	1.29	12.12
	Bates-Granger (2)	1.29	0.86	67.16	2.05	1.26	112.66	0.98	0.81	287.11	1.12	0.95	346.00	0.79	0.65	6.43	1.60	1.27	11.98
	Bates-Granger (3)	1.32	0.88	68.71	2.07	1.28	115.76	1.03	0.85	305.01	1.14	0.96	355.75	0.81	0.67	6.58	1.60	1.29	12.14
	Bates-Granger (4)	1.28	0.85	65.57	2.06	1.27	113.69	1.01	0.84	297.02	1.11	0.93	348.88	0.79	0.64	6.37	1.55	1.23	11.51
	Bates-Granger (5)	1.29	0.87	67.29	2.06	1.28	114.97	0.99	0.83	291.72	1.12	0.95	348.74	0.81	0.66	6.50	1.60	1.30	12.25
	Granger-Ramanathan (1)	2.02	1.51	116.43	2.93	2.59	225.72	1.58	1.38	287.80	1.53	1.23	180.08	1.19	0.98	9.64	2.85	2.30	22.05
	Granger-Ramanathan (2)	2.21	1.59	122.78	2.60	2.15	209.95	1.51	1.26	284.63	1.46	1.14	220.66	1.23	0.91	8.78	2.87	2.28	21.72
	Granger-Ramanathan (3)	2.13	1.63	139.29	2.76	2.34	211.44	1.44	1.27	287.85	1.99	1.37	186.90	1.16	0.92	8.93	2.56	2.18	20.51
	AFTER	1.29	0.88	65.00	2.05	1.26	110.15	1.02	0.85	298.76	1.13	0.95	353.34	0.81	0.67	6.60	1.55	1.26	11.89
	Median Forecast	1.32	0.87	68.24	2.07	1.28	116.52	1.03	0.86	304.50	1.15	0.97	360.49	0.82	0.67	6.61	1.59	1.29	12.13
	Trimmed Mean Forecast	1.32	0.88	68.69	2.07	1.28	115.91	1.03	0.85	305.50	1.14	0.96	356.35	0.81	0.67	6.58	1.60	1.29	12.15
	PEW	1.33	0.98	73.23	2.00	1.42	110.08	1.11	0.93	313.50	1.18	0.85	406.75	0.84	0.69	6.83	1.54	1.41	13.43
Factor An.	Principal Component Forecast	1.35	0.95	70.79	1.93	1.48	101.77	1.11	0.89	327.49	1.15	0.99	340.20	0.85	0.74	7.34	1.64	1.47	14.12
	Principal Component Forecast (AIC)	1.15	0.82	57.72	2.00	1.51	103.11	1.17	0.97	332.11	1.14	0.96	265.59	0.86	0.74	7.32	1.67	1.49	14.29
	Principal Component Forecast (BIC)	1.23	0.86	60.47	1.95	1.49	102.65	1.16	0.96	330.97	1.17	1.00	277.45	0.88	0.77	7.67	1.64	1.47	14.12
Shrinkage	Empirical Bayes Estimator	1.92	1.44	114.37	2.28	1.85	174.89	1.25	1.10	275.73	1.68	1.21	226.84	0.98	0.77	7.56	2.28	1.91	18.02
	Kappa-Shrinkage	1.24	0.84	62.61	2.07	1.61	132.50	0.94	0.79	234.90	0.97	0.83	277.38	0.91	0.74	7.34	1.76	1.37	13.06
	Two-Step Egalitarian LASSO	1.30	0.90	65.04	2.02	1.54	122.16	1.22	1.00	363.66	1.34	1.11	325.17	0.85	0.69	6.84	1.72	1.45	13.86
BMA	BMA (Marginal Likelihood)	1.45	1.04	75.40	1.93	1.40	101.77	1.23	1.04	374.66	1.13	0.96	352.05	0.84	0.70	6.92	1.56	1.43	13.55
	BMA (Predictive Likelihood)	1.99	1.51	143.15	2.57	1.97	197.58	1.74	1.39	616.53	1.37	1.08	465.58	1.42	1.11	10.61	3.47	2.91	26.97
Alternative	ANN	1.41	1.02	84.51	2.06	1.53	104.87	1.08	0.90	289.06	1.04	0.81	302.34	0.90	0.75	7.36	1.42	1.18	11.12
	EP-NN	2.95	1.81	191.07	11.68	6.16	587.12	3.13	2.02	497.31	5.12	2.53	375.98	1.88	1.37	13.04	3.30	2.23	21.27
	Bagging	15.15	6.52	528.38	2.52	2.06	136.01	1.04	0.86	204.62	1.36	1.07	204.78	2.05	1.89	18.31	3.69	3.34	31.24
	Componentwise Boosting	1.48	0.96	64.75	1.93	1.42	102.74	1.19	0.99	392.98	1.14	0.98	348.37	1.67	1.49	14.04	1.81	1.61	14.94
	AdaBoost	1.38	1.08	80.94	2.31	1.95	148.02	1.10	0.90	290.48	0.87	0.70	158.72	1.49	1.31	12.35	1.85	1.55	14.23
APM	c-APM (Constant)	1.32	0.86	62.31	2.09	1.30	115.71	1.02	0.85	300.74	1.13	0.98	350.77	0.79	0.64	6.31	1.55	1.26	11.91
	c-APM (Q-learning)	1.45	0.95	66.38	2.15	1.39	127.08	1.03	0.85	282.16	1.16	1.02	330.04	0.82	0.65	6.43	1.70	1.30	12.14
	Market for Kernels	1.27	0.85	63.38	2.00	1.22	103.97	1.03	0.88	293.17	1.09	0.90	354.08	0.83	0.71	7.02	1.63	1.34	12.58
	Best Individual	1.15	0.81	59.66	1.95	1.20	99.44	0.97	0.77	256.27	1.06	0.88	299.56	0.79	0.63	6.17	1.46	1.06	9.65
	Median Individual	1.37	0.92	72.63	2.08	1.30	118.81	1.06	0.88	305.85	1.16	0.98	360.47	0.85	0.68	6.73	1.64	1.37	12.83
	Worst Individual	1.64	1.12	106.49	2.31	1.55	148.26	1.26	1.06	382.87	1.37	1.17	442.27	1.05	0.84	8.29	1.97	1.71	16.14

Table 5.3: Performance of forecast combinations of ECB SPF forecasts using the training window of the length: 45

Class	Forecast Combination Method	RGDP						HICP						UNEM					
		1Y			2Y			1Y			2Y			1Y			2Y		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Simple	Equal Weights	0.69	0.54	55.70	1.25	0.87	127.59	0.90	0.79	281.33	1.15	0.95	456.90	0.67	0.57	5.25	1.41	1.19	10.75
	Bates-Granger (1)	0.69	0.54	55.13	1.25	0.86	127.26	0.90	0.79	280.03	1.15	0.95	457.09	0.67	0.57	5.25	1.40	1.18	10.68
	Bates-Granger (2)	0.69	0.54	53.17	1.24	0.87	126.86	0.88	0.78	267.35	1.15	0.96	455.16	0.69	0.59	5.43	1.35	1.12	10.13
	Bates-Granger (3)	0.69	0.54	55.47	1.25	0.87	127.46	0.90	0.79	280.81	1.15	0.95	456.98	0.67	0.57	5.25	1.41	1.19	10.72
	Bates-Granger (4)	0.66	0.52	52.59	1.23	0.85	124.58	0.88	0.77	269.85	1.12	0.92	447.66	0.65	0.54	5.03	1.34	1.10	9.82
	Bates-Granger (5)	0.68	0.53	53.58	1.25	0.85	125.83	0.88	0.77	267.93	1.13	0.93	447.16	0.67	0.57	5.24	1.39	1.18	10.63
	Granger-Ramanathan (1)	1.28	1.01	99.95	2.13	1.84	208.15	1.24	1.00	233.41	1.21	1.03	353.85	1.03	0.86	8.00	2.01	1.65	15.45
	Granger-Ramanathan (2)	1.25	1.03	102.63	2.14	1.68	193.93	1.25	0.99	236.27	1.20	1.04	378.02	1.01	0.78	7.19	1.87	1.56	14.35
	Granger-Ramanathan (3)	1.30	1.05	105.58	2.20	1.74	213.79	1.22	1.01	258.32	1.25	1.03	328.87	1.03	0.77	7.16	1.97	1.69	15.77
	AFTER	0.70	0.56	52.38	1.20	0.84	121.37	0.90	0.79	274.25	1.15	0.95	455.59	0.68	0.57	5.29	1.33	1.12	10.14
	Median Forecast	0.68	0.53	53.91	1.26	0.88	129.27	0.90	0.79	279.72	1.16	0.96	463.44	0.68	0.57	5.30	1.40	1.19	10.80
	Trimmed Mean Forecast	0.69	0.54	54.92	1.25	0.87	127.59	0.90	0.79	281.33	1.15	0.95	457.92	0.67	0.57	5.25	1.41	1.19	10.75
	PEW	0.69	0.53	46.52	1.15	0.96	110.89	0.96	0.81	303.14	1.33	1.00	554.58	0.73	0.61	5.66	1.47	1.27	11.66
Factor An.	Principal Component Forecast	0.67	0.53	52.81	1.04	0.91	93.56	0.91	0.78	293.26	1.17	0.96	450.30	0.69	0.61	5.63	1.42	1.32	12.30
	Principal Component Forecast (AIC)	0.69	0.52	42.98	1.04	0.91	93.56	0.97	0.83	295.82	1.22	1.04	431.21	0.69	0.57	5.21	1.35	1.22	11.21
	Principal Component Forecast (BIC)	0.71	0.53	45.29	1.04	0.91	93.56	0.94	0.80	294.28	1.20	1.04	454.65	0.68	0.59	5.41	1.42	1.32	12.30
Shrinkage	Empirical Bayes Estimator	1.10	0.89	88.61	1.75	1.29	171.09	1.07	0.90	231.43	1.08	0.88	288.66	0.92	0.70	6.45	1.72	1.43	13.24
	Kappa-Shrinkage	0.85	0.68	64.03	1.67	1.40	163.51	0.95	0.79	168.40	1.01	0.83	300.11	0.88	0.74	6.88	1.47	1.09	10.20
	Two-Step Egalitarian LASSO	0.84	0.70	64.46	0.78	0.55	72.88	0.96	0.79	291.57	1.29	1.09	495.66	0.72	0.58	5.35	1.14	0.96	8.68
BMA	BMA (Marginal Likelihood)	0.78	0.60	50.99	1.08	0.91	102.18	1.05	0.87	359.91	1.27	1.02	517.06	0.77	0.64	5.89	1.48	1.27	11.50
	BMA (Predictive Likelihood)	1.32	1.10	133.75	1.80	1.51	193.76	1.47	1.13	498.99	2.03	1.68	657.95	1.82	1.48	13.10	3.26	2.84	25.70
Alternative	ANN	1.09	0.79	90.51	1.03	0.86	96.41	1.23	1.06	371.49	1.22	0.99	454.39	0.79	0.66	6.18	1.40	1.17	10.70
	EP-NN	1.76	1.41	168.02	3.04	2.23	329.68	1.56	1.13	351.82	1.93	1.58	680.42	1.61	1.07	9.88	1.77	1.48	13.86
	Bagging	0.98	0.74	62.47	1.23	1.09	105.10	0.96	0.84	195.23	0.87	0.78	284.67	3.61	3.49	32.65	3.75	3.40	30.48
	Componentwise Boosting	0.79	0.59	60.31	1.08	0.91	100.98	1.19	0.98	433.54	1.20	0.96	486.25	1.86	1.61	14.48	1.94	1.63	14.40
	AdaBoost	0.90	0.70	53.61	1.27	1.14	109.86	1.04	0.82	213.79	0.95	0.78	223.49	1.45	1.24	11.14	1.94	1.51	13.32
APM	c-APM (Constant)	0.69	0.51	46.93	1.26	0.87	125.48	0.91	0.81	287.79	1.16	0.98	452.44	0.72	0.60	5.60	1.34	1.12	10.14
	c-APM (Q-learning)	0.66	0.50	50.51	1.48	1.07	151.72	0.88	0.75	281.64	1.19	0.99	478.62	0.73	0.61	5.60	1.15	0.92	8.38
	Market for Kernels	0.73	0.57	51.11	1.16	0.83	116.89	0.89	0.77	253.65	1.10	0.90	443.45	0.73	0.61	5.66	1.32	1.17	10.67
	Best Individual	0.61	0.47	42.10	1.05	0.77	104.11	0.71	0.62	195.21	1.03	0.87	375.21	0.61	0.47	4.40	1.25	0.99	8.57
	Median Individual	0.73	0.58	57.42	1.27	0.92	132.08	0.93	0.81	275.11	1.16	0.99	467.62	0.72	0.58	5.38	1.48	1.25	11.34
	Worst Individual	1.01	0.77	98.38	1.49	1.08	163.82	1.29	1.09	405.25	1.41	1.22	579.66	0.82	0.72	6.69	1.84	1.62	14.79

Table 5.4: P-values from the DM test of equal forecast accuracy: equal weights against the remaining combinations of forecasts from the ECB SPF

Class	Forecast Combination Method	w = 25						w = 35						w = 45					
		RGDP		HICP		UNEM		RGDP		HICP		UNEM		RGDP		HICP		UNEM	
		1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y
Simple	Equal Weights	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Bates-Granger (1)	0.14	0.02^{**}	0.10^{*}	0.15	0.44	0.20	0.15	0.05^{*}	0.17	0.29	0.34	0.06^{*}	0.73	0.15	0.35	0.82	0.59	0.09^{*}
	Bates-Granger (2)	0.03^{**}	0.14	0.00^{***}	0.01^{***}	0.04^{**}	0.04^{**}	0.17	0.11	0.02^{**}	0.11	0.10^{*}	0.97	0.84	0.38	0.15	0.74	$0.05^{\dagger}$	0.01^{**}
	Bates-Granger (3)	0.14	0.02^{**}	0.10^{*}	0.14	0.41	0.19	0.15	0.05^{*}	0.17	0.29	0.33	0.06^{*}	0.72	0.15	0.35	0.82	0.61	0.08^{*}
	Bates-Granger (4)	0.05^{**}	0.00^{***}	0.01^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.06^{*}	0.01^{**}	0.01^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.05^{*}	0.06^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (5)	0.13	0.57	0.03^{**}	0.01^{***}	0.18	0.29	0.11	0.66	0.02^{**}	0.02^{**}	0.58	0.66	0.37	0.98	0.05^{**}	0.04^{**}	0.81	0.27
	Granger-Ramanathan (1)	$0.04^{\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.01^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.06^{\dagger}$	$0.06^{\dagger}$	$0.07^{\dagger}$	$0.07^{\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.21	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.11	0.77	$0.00^{\dagger\dagger\dagger}$	$0.09^{\dagger}$
	Granger-Ramanathan (2)	$0.03^{\dagger\dagger}$	$0.01^{\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger}$	$0.03^{\dagger\dagger}$	$0.09^{\dagger}$	0.21	$0.01^{\dagger\dagger}$	0.31	$0.01^{\dagger\dagger}$	$0.01^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.06^{\dagger}$	0.83	$0.01^{\dagger\dagger}$	$0.07^{\dagger}$
	Granger-Ramanathan (3)	$0.02^{\dagger\dagger}$	$0.05^{\dagger}$	$0.02^{\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.07^{\dagger}$	$0.05^{\dagger\dagger}$	0.15	$0.01^{\dagger\dagger\dagger}$	0.14	$0.01^{\dagger\dagger\dagger}$	$0.03^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.11	0.63	$0.01^{\dagger\dagger}$	$0.03^{\dagger\dagger}$
	AFTER	0.33	0.08^{*}	0.20	0.11	0.68	0.14	0.39	0.28	0.32	0.11	0.52	0.07^{*}	0.81	0.18	0.44	0.41	0.41	0.08^{*}
	Median Forecast	0.22	0.70	0.61	0.62	0.28	0.77	0.38	0.92	0.74	0.19	0.28	0.47	0.42	$0.02^{\dagger\dagger}$	0.79	0.28	$0.08^{\dagger}$	0.67
	Trimmed Mean Forecast	0.97	0.23	0.94	0.41	0.72	0.74	0.93	0.40	0.45	0.37	0.72	0.66	0.22	0.15	0.42	0.87	0.45	0.75
	PEW	0.27	0.34	0.41	0.96	0.44	0.79	0.96	0.74	0.39	0.82	0.65	0.77	0.99	0.68	0.68	0.33	0.13	0.61
Factor An.	Principal Component Forecast	0.16	0.50	0.79	0.44	0.56	0.89	0.54	0.61	0.36	0.90	0.29	0.84	0.51	0.56	0.93	0.79	0.75	0.98
	Principal Component Forecast (AIC)	0.65	0.46	0.37	0.66	$0.09^{\dagger}$	0.70	0.36	0.78	$0.08^{\dagger}$	1.00	0.38	0.74	1.00	0.56	0.23	0.46	0.82	0.82
	Principal Component Forecast (BIC)	0.43	0.52	0.44	0.69	0.13	0.85	0.47	0.66	$0.10^{\dagger}$	0.84	0.27	0.84	0.91	0.56	0.45	0.44	0.83	0.98
Shrinkage	Empirical Bayes Estimator	$0.03^{\dagger\dagger}$	$0.07^{\dagger}$	$0.02^{\dagger\dagger}$	$0.03^{\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.06^{\dagger}$	$0.08^{\dagger}$	0.59	$0.07^{\dagger}$	0.28	$0.01^{\dagger\dagger\dagger}$	0.11	$0.01^{\dagger\dagger\dagger}$	$0.07^{\dagger}$	0.24	0.68	$0.03^{\dagger\dagger}$	0.12
	Kappa-Shrinkage	0.55	0.22	0.33	0.36	0.82	0.70	0.27	0.99	0.22	0.02^{**}	$0.00^{\dagger\dagger\dagger}$	0.20	$0.01^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.69	0.28	$0.00^{\dagger\dagger\dagger}$	0.83
	Two-Step Egalitarian LASSO	$0.02^{\dagger\dagger}$	0.52	0.81	0.86	0.13	0.16	0.75	0.86	$0.00^{\dagger\dagger\dagger}$	$0.07^{\dagger}$	$0.04^{\dagger\dagger}$	0.31	0.18	0.11	0.63	$0.06^{\dagger}$	0.37	0.06^{*}
BMA	BMA (Marginal Likelihood)	0.14	0.20	0.96	0.24	0.25	0.73	0.45	0.55	$0.03^{\dagger\dagger}$	0.90	0.72	0.84	0.59	0.55	0.39	0.35	$0.07^{\dagger}$	0.63
	BMA (Predictive Likelihood)	$0.02^{\dagger\dagger}$	0.24	0.11	0.87	0.14	0.11	$0.01^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.08^{\dagger}$	0.25	$0.02^{\dagger\dagger}$	$0.06^{\dagger}$	$0.02^{\dagger\dagger}$	$0.08^{\dagger}$	$0.08^{\dagger}$	$0.04^{\dagger\dagger}$	$0.03^{\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$
Alternative	ANN	0.16	0.32	0.96	0.67	$0.03^{\dagger\dagger}$	0.37	0.65	0.96	0.61	0.32	$0.04^{\dagger\dagger}$	0.36	$0.10^{\dagger}$	0.47	$0.01^{\dagger\dagger\dagger}$	0.44	$0.01^{\dagger\dagger\dagger}$	0.93
	EP-NN	$0.03^{\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.29	$0.01^{\dagger\dagger\dagger}$	$0.04^{\dagger\dagger}$	0.17	0.18	$0.05^{\dagger\dagger}$	0.15	$0.01^{\dagger\dagger\dagger}$	$0.10^{\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.02^{\dagger\dagger}$	0.15	$0.02^{\dagger\dagger}$	0.10	0.29
	Bagging	$0.00^{\dagger\dagger\dagger}$	0.84	$0.02^{\dagger\dagger}$	0.16	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.20	0.30	0.96	0.44	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.22	0.97	0.67	0.15	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$
	Componentwise Boosting	$0.10^{\dagger}$	0.28	0.91	0.24	$0.00^{\dagger\dagger\dagger}$	0.79	0.21	0.54	0.16	0.98	$0.01^{\dagger\dagger\dagger}$	0.44	0.36	0.54	0.13	0.39	$0.01^{\dagger\dagger\dagger}$	$0.03^{\dagger\dagger}$
	AdaBoost	0.68	0.84	0.89	0.04^{**}	$0.01^{\dagger\dagger\dagger}$	0.49	0.72	0.60	0.24	0.16	$0.01^{\dagger\dagger\dagger}$	0.39	0.29	0.95	0.13	0.39	$0.01^{\dagger\dagger}$	$0.06^{\dagger}$
APM	c-APM (Constant)	0.96	0.42	0.18	0.37	0.71	0.11	0.94	0.42	0.27	0.58	0.42	0.06^{*}	0.99	0.69	0.42	0.63	$0.02^{\dagger\dagger}$	0.12
	c-APM (Q-learning)	0.79	0.02^{**}	0.31	0.54	0.43	0.45	0.22	$0.03^{\dagger\dagger}$	0.93	0.78	0.84	0.25	0.36	$0.06^{\dagger}$	0.35	0.41	0.14	0.02^{**}
	Market for Kernels	0.23	0.03^{**}	0.74	0.00^{***}	1.00	0.93	0.61	0.09^{*}	0.94	0.15	0.68	0.55	0.71	0.21	0.70	0.06^{*}	0.12	0.41

*The forecasts are significantly better than forecasts from the equal weights

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

†The forecasts are significantly worse than forecasts from the equal weights

$^{\dagger\dagger\dagger}p < 0.01$, $^{\dagger\dagger}p < 0.05$, $^{\dagger}p < 0.1$

Table 5.5: P-values from the DM test of equal forecast accuracy: Market for Kernels against the remaining combinations of forecasts from the ECB SPF

Class	Forecast Combination Method	w = 25						w = 35						w = 45					
		RGDP		HICP		UNEM		RGDP		HICP		UNEM		RGDP		HICP		UNEM	
		1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y	1Y	2Y
Simple	Equal Weights	0.23	0.03**	0.74	0.00***	1.00	0.93	0.61	0.09*	0.94	0.15	0.68	0.55	0.71	0.21	0.70	0.06*	0.12	0.41
	Bates-Granger (1)	0.24	0.03**	0.81	0.00***	0.96	1.00	0.67	0.10*	0.89	0.15	0.63	0.40	0.69	0.21	0.72	0.06*	0.11	0.45
	Bates-Granger (2)	0.37	0.03**	0.55	0.03**	0.68	0.87	0.83	0.21	0.04††	0.50	0.41	0.51	0.69	0.18	0.85	0.05**	0.24	0.74
	Bates-Granger (3)	0.23	0.03**	0.76	0.00***	0.98	0.96	0.63	0.10*	0.92	0.15	0.66	0.48	0.70	0.21	0.71	0.06*	0.12	0.43
	Bates-Granger (4)	0.37	0.05**	0.98	0.01***	0.51	0.43	0.93	0.13	0.59	0.52	0.27	0.04††	0.49	0.26	0.78	0.34	0.05††	0.87
	Bates-Granger (5)	0.32	0.03**	0.56	0.01**	0.81	0.88	0.83	0.13	0.18	0.40	0.62	0.36	0.57	0.25	0.72	0.21	0.13	0.48
	Granger-Ramanathan (1)	0.04**	0.02**	0.01***	0.00***	0.07*	0.06*	0.04**	0.04**	0.00***	0.18	0.00***	0.01***	0.01**	0.00***	0.11	0.56	0.00***	0.03**
	Granger-Ramanathan (2)	0.02**	0.01**	0.01***	0.00***	0.02**	0.02**	0.08*	0.16	0.01**	0.26	0.01**	0.01**	0.00***	0.00***	0.06*	0.61	0.02**	0.01**
	Granger-Ramanathan (3)	0.02**	0.04**	0.01**	0.02**	0.03**	0.06*	0.03**	0.12	0.01***	0.13	0.00***	0.02**	0.02**	0.01***	0.12	0.47	0.04**	0.00***
	AFTER	0.25	0.05*	0.93	0.00***	0.95	0.52	0.82	0.12	0.72	0.21	0.58	0.04††	0.64	0.30	0.79	0.06*	0.10	0.87
	Median Forecast	0.28	0.04**	0.71	0.00***	0.87	0.99	0.67	0.11	0.96	0.07*	0.84	0.38	0.62	0.17	0.73	0.03**	0.17	0.41
	Trimmed Mean Forecast	0.23	0.03**	0.74	0.00***	1.00	0.93	0.61	0.09*	0.94	0.14	0.68	0.55	0.67	0.21	0.70	0.05**	0.12	0.41
	PEW	0.19	0.68	0.42	0.55	0.55	0.78	0.56	0.97	0.48	0.53	0.91	0.69	0.55	0.97	0.66	0.20	0.95	0.24
Factor An.	Principal Component Forecast	0.04**	0.78	0.71	0.91	0.61	0.84	0.36	0.76	0.52	0.35	0.74	0.95	0.45	0.70	0.83	0.30	0.38	0.52
	Principal Component Forecast (AIC)	0.22	0.78	0.25	0.95	0.08*	0.63	0.28	0.98	0.18	0.70	0.54	0.84	0.57	0.70	0.31	0.12	0.43	0.82
	Principal Component Forecast (BIC)	0.17	0.86	0.36	0.88	0.14	0.81	0.44	0.82	0.21	0.54	0.42	0.95	0.72	0.70	0.50	0.07*	0.29	0.52
Shrinkage	Empirical Bayes Estimator	0.03**	0.06*	0.02**	0.02**	0.03**	0.06*	0.05**	0.49	0.05*	0.25	0.01***	0.11	0.06*	0.08*	0.25	0.92	0.12	0.04**
	Kappa-Shrinkage	0.23	0.03**	0.74	0.00***	1.00	0.93	0.62	0.79	0.14	0.12	0.13	0.23	0.29	0.00***	0.61	0.46	0.01**	0.45
	Two-Step Egalitarian LASSO	0.02**	0.91	0.99	0.68	0.12	0.16	0.61	0.94	0.04**	0.01**	0.71	0.48	0.10	0.09†	0.61	0.03**	0.72	0.01†††
BMA	BMA (Marginal Likelihood)	0.12	0.36	0.84	0.52	0.25	0.68	0.18	0.73	0.06*	0.73	0.91	0.76	0.48	0.73	0.41	0.16	0.48	0.35
	BMA (Predictive Likelihood)	0.01***	0.15	0.12	0.57	0.14	0.11	0.01***	0.00***	0.08*	0.13	0.02**	0.06*	0.03**	0.06*	0.10*	0.03**	0.03**	0.00***
Alternative	ANN	0.12	0.15	0.88	0.36	0.12	0.30	0.29	0.78	0.62	0.54	0.23	0.32	0.23	0.60	0.01**	0.10	0.37	0.60
	EP-NN	0.03**	0.02**	0.00***	0.29	0.01***	0.04**	0.16	0.17	0.05*	0.15	0.01***	0.10*	0.00***	0.02**	0.16	0.02**	0.12	0.11
	Bagging	0.00***	0.98	0.01**	0.13	0.00***	0.00***	0.20	0.21	0.98	0.37	0.00***	0.00***	0.22	0.84	0.48	0.23	0.00***	0.00***
	Componentwise Boosting	0.07*	0.47	0.80	0.54	0.00***	0.77	0.19	0.72	0.20	0.57	0.01***	0.54	0.34	0.72	0.16	0.05*	0.01***	0.06*
	AdaBoost	0.08*	0.64	0.94	0.10	0.01***	0.50	0.29	0.49	0.26	0.26	0.01***	0.48	0.17	0.74	0.13	0.51	0.02**	0.10
APM	c-APM (Constant)	0.30	0.08*	0.71	0.02**	0.87	0.58	0.64	0.11	0.56	0.46	0.27	0.04††	0.73	0.24	0.58	0.04**	0.75	0.80
	c-APM (Q-learning)	0.16	0.25	0.24	0.01**	0.63	0.58	0.28	0.04**	0.89	0.46	0.82	0.53	0.41	0.09*	0.88	0.14	0.95	0.00†††
	Market for Kernels	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

*The Market for Kernels forecasts are significantly better than the given forecasts
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$
†The Market for Kernels forecasts are significantly worse than the given forecasts
†††$p < 0.01$, ††$p < 0.05$, †$p < 0.1$

5.4 Forecasting U.S. Treasury Futures Volatility

As in the ECB SPF application, the forecast combinations methods described in the chapter 3 were estimated (trained) on a rolling windows of lengths 100, 200 and 500 and then the 1-step-ahead out-of-sample forecast combinations of the U.S. Treasury futures log-returns realized volatility forecasts were obtained. The figure 5.4 shows the best performing forecast combinations from each class on the TU (2 year) futures datasets. These include combinations of 1,5 and 22-steps-ahead individual realized volatility forecasts trained on rolling windows of length 100 and 500. The outputs of forecasts combinations trained on rolling windows of length 200 are visually very similar to those trained on rolling windows of length 100 and so we do not display it for space reasons. The definition of the best in class method used is the same as in the ECB SPF application results from the previous section. The displayed part of samples was cut off bellow the 12th of August of 2009, so that they can be aligned and visually compared. The figure reflects that the individual forecasts of the realized volatility from the volatility models presented in the section 4.2.2, especially the HAR and VAR models, are already so accurate, that within each class of combination methods there was atleast one method that was able to filter out the inferior individual forecasts given by e.g. the Historical Volatility. Disregarding the length of the rolling window, the forecasts overlay very well with the true realized volatility, except the occasional sudden jumps for the one-step-ahead individual forecasts (h = 1) and become substantially more blur as the forecast horizon and so the inaccuracy of the combined individual forecasts increases (h=5, h=22). This implies that there still is some non-negligible difference in how the different forecast combinations assign weights. The figures A.1, A.2 and A.3 show combinations of forecasts of realized volatility of the FV, TY and US futures respectively. These figures deliver very similar message as the TU figure and are therefore presented in the appendix A.

The table 5.6 summarizes the forecast accuracy of the combination methods on the realized volatility of the TU (2 Year) U.S. Treasury futures log-returns datasets. The forecast performance across all methods deteriorates with the increasing h (forecast horizon of the individual forecasts) in terms of both RMSE and MAPE. On the other hand, it notably improves with the increasing w (the length of the training window). As in the ECB SPF application, the forecasts combinations most of the methods from the simple, factor analytic and APM classes generally achieve very decent results in terms of the accuracy measures in both shorter and longer training windows. With the increasing length of the training window, the OLS based Granger-Ramanathan methods radically improve and catch up with the other combining methods. Moreover, the outsiders from the ECB SPF application, the BMA with the weights based on the predictive likelihood and the artificial neural networks

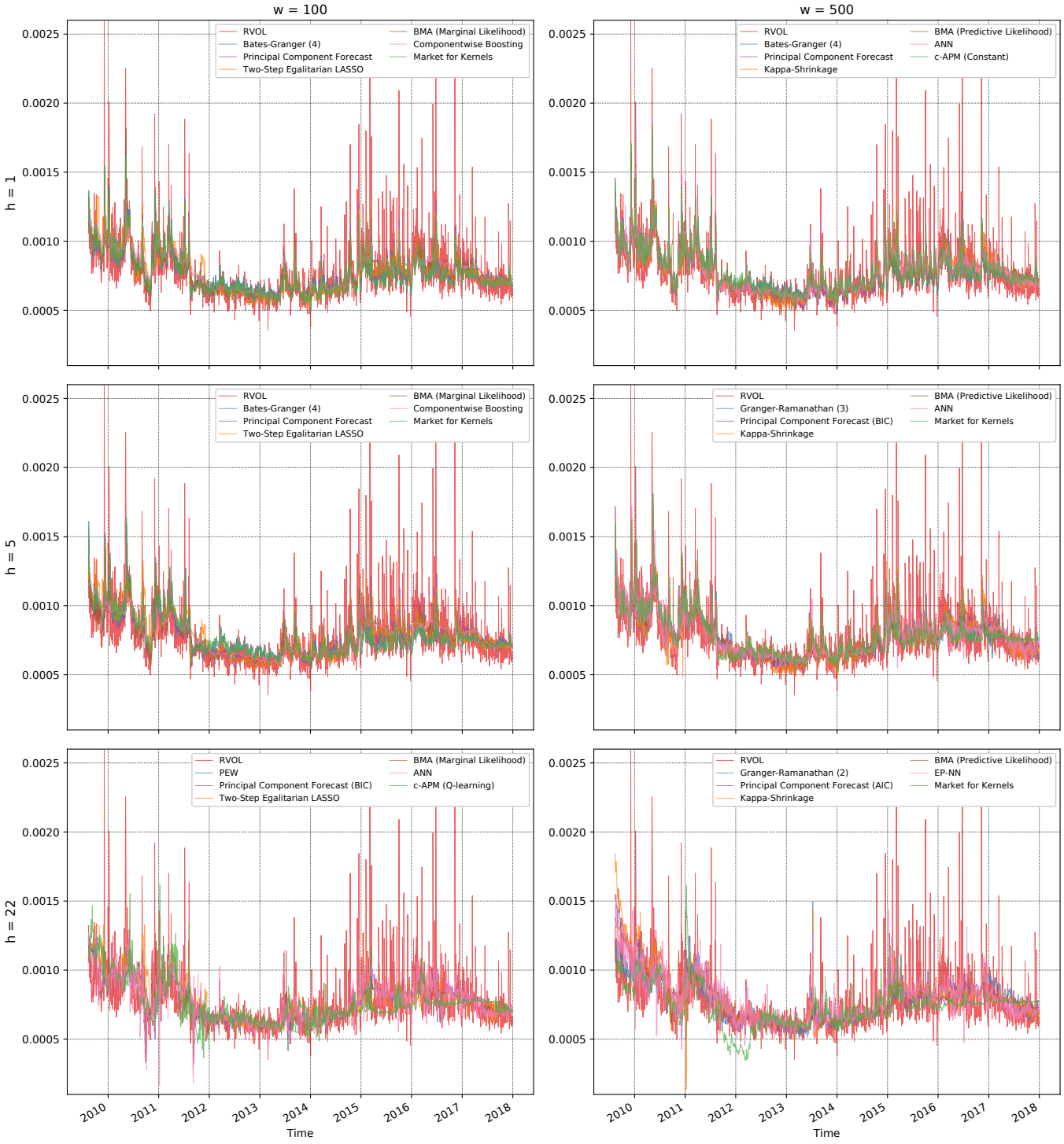

Figure 5.4: Best combinations of h-steps-ahead forecasts of realized volatility of TU (2 Year) U.S. Treasury futures log-returns, trained on a rolling window of length w

(ANN and EP-NN) start to shine with the increasing sample length and growing uncertainty of the individual realized volatility forecasts. For h=22 and w=500 they even outperform the best individual forecast. Overall, most of the methods perform relatively comparably with the median forecast individual in short samples and tend to improve down to the best individual forecast in longer samples. As with the figures, the inference from the tables of forecast combination accuracy measures in cases of FV, TY and US futures (tables A.1, A.2 and A.3) is the same in the case of TU and hence these tables are presented only in the appendix A.

The third hypothesis tested in this work is that the equal weights (simple average) combination method forecasts of the realized volatility of the U.S. Treasury futures log-returns are equally accurate as the forecasts of the other described forecast combinations. The tables 5.7 and 5.8 show the p-values from the DM-test for the US, FV and TY, US futures respectively. We can reject the null hypothesis in favour of the Bates-Granger in vast majority of the datasets. We can also see that the Granger-Ramanathan methods significantly underperform the equal weights in short sample and significantly outperform the equal weights in large samples and long horizons. The factor analytic combination forecasts tend to significantly outperform the equal weights forecasts, especially in the longer horizons. The BMA methods forecasts as well as the ANN methods forecasts also significantly outperform the equals weights forecasts in the longer horizons. The inference about APM is mixed. While Market for Kernels tends to outperform the equal weights in majority of cases, the c-APM (especially the constant betting function version) significantly underperforms the equals weights on numerous TY and US futures datasets. The overall worst performing forecast combination methods are found to be the Empirical Bayes Estimator and the Bagging method, for which the null hypothesis is rejected in favour of equal weights in almost every case.

Our fourth hypothesis is that the newly proposed method, the Market for Kernels, forecasts of the realized volatility of the U.S. Treasury futures log-returns are equally accurate as the forecasts of the other described forecast combinations. The tables 5.9 and 5.10 summarize the DM-test p-values for the US, FV and TY, US futures respectively. The Market for Kernels is found to either significantly outperform or show a performance statistically indifferent from most of the combination methods on most of the futures datasets. The exceptions being the factor analytic, BMA and artifical neural networks methods in large samples and long horizons. The most striking exception is the Bates-Granger (4) method, for which the hypothesis can be rejected in favour on very low significance levels on majority of the datasets.

The last hypothesis in this study is that the Market for Kernels forecasts of the realized volatility of the U.S. Treasury futures log-returns are equally accurate as the individual forecasts it combines. It shows that we can reject the hypothesis in favour

Table 5.6: Performance of forecast combinations, trained on a rolling window of length w, of inidividual h-steps-ahead forecasts of realized volatility of log-returns of U.S. Treasury futures: TU (2 Year)

Class	Forecast Combination Method	h = 1						h = 5						h = 22					
		w = 100		w = 200		w = 500		w = 100		w = 200		w = 500		w = 100		w = 200		w = 500	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Simple	Equal Weights	2.71	12.86	2.50	12.56	1.81	12.17	3.13	15.02	2.83	14.67	1.94	14.04	3.92	21.09	3.62	20.79	2.31	20.03
	Bates-Granger (1)	2.65	12.94	2.44	12.58	1.74	12.01	3.08	15.28	2.79	14.96	1.87	14.07	3.75	19.16	3.50	19.12	2.16	18.39
	Bates-Granger (2)	2.67	12.63	2.44	12.23	1.74	11.84	3.11	14.70	2.80	14.36	1.87	13.56	3.94	21.53	3.64	21.45	2.31	21.09
	Bates-Granger (3)	2.68	12.60	2.46	12.28	1.77	11.79	3.10	14.89	2.81	14.55	1.90	13.80	3.85	20.16	3.56	19.97	2.24	19.24
	Bates-Granger (4)	2.63	12.87	2.42	12.57	1.73	12.13	3.03	14.58	2.74	14.25	1.83	13.62	3.58	16.42	3.33	16.13	1.92	15.24
	Bates-Granger (5)	2.71	13.14	2.51	12.87	1.82	12.48	3.15	15.08	2.85	14.73	1.94	14.07	3.93	21.30	3.62	21.00	2.31	20.27
	Granger-Ramanathan (1)	3.00	15.04	2.55	13.65	1.74	12.69	3.48	16.39	2.97	15.74	1.90	14.28	3.59	17.33	3.47	17.62	2.08	15.79
	Granger-Ramanathan (2)	2.99	14.69	2.56	13.24	1.73	12.45	3.60	16.63	2.97	15.22	1.87	13.69	4.56	17.93	3.78	16.79	1.89	14.64
	Granger-Ramanathan (3)	3.02	15.51	2.59	13.73	1.73	12.47	3.50	16.98	3.04	15.50	1.82	13.55	3.61	17.10	3.30	16.28	1.93	15.35
	AFTER	2.71	12.86	2.50	12.56	1.81	12.17	3.13	15.02	2.83	14.67	1.94	14.04	3.92	21.09	3.62	20.79	2.31	20.03
	Median Forecast	2.64	13.42	2.43	13.13	1.75	12.73	3.13	16.31	2.83	16.02	1.92	15.50	4.13	24.89	3.76	24.56	2.48	24.35
	Trimmed Mean Forecast	2.71	12.86	2.50	12.56	1.81	12.17	3.13	15.02	2.83	14.67	1.94	14.04	3.92	21.09	3.62	20.79	2.31	20.03
	PEW	2.63	13.43	2.45	13.17	1.74	12.49	2.94	14.79	2.76	14.89	1.92	14.63	3.28	16.10	3.25	17.63	2.45	18.88
Factor An.	Principal Component Forecast	2.61	13.16	2.43	12.94	1.75	12.63	2.99	14.57	2.78	14.41	1.88	14.22	3.45	16.15	3.39	16.70	2.16	17.92
	Principal Component Forecast (AIC)	2.65	13.55	2.46	13.33	1.74	12.92	3.01	14.90	2.79	14.85	1.84	14.08	3.11	15.62	2.91	16.05	2.02	15.53
	Principal Component Forecast (BIC)	2.66	13.27	2.45	13.13	1.74	12.96	3.01	14.69	2.80	14.84	1.85	14.02	3.08	15.26	2.89	15.92	2.02	15.66
Shrinkage	Empirical Bayes Estimator	6.73	30.58	4.98	23.77	2.34	16.27	9.47	33.96	4.49	21.85	2.32	16.57	16.85	41.74	5.23	31.51	3.25	25.62
	Kappa-Shrinkage	2.94	14.66	2.54	13.53	1.74	12.65	3.39	15.93	2.95	15.58	1.89	14.23	3.47	16.73	3.43	17.41	2.08	15.76
	Two-Step Egalitarian LASSO	2.72	14.21	2.52	14.80	1.83	14.27	3.07	15.92	2.82	16.24	1.91	14.62	3.37	17.06	3.30	18.97	2.21	18.62
BMA	BMA (Marginal Likelihood)	2.65	13.59	2.44	13.34	2.00	14.69	2.99	14.66	2.83	15.16	1.92	14.56	3.12	14.92	3.14	16.18	2.03	15.77
	BMA (Predictive Likelihood)	2.74	13.86	2.47	13.30	1.73	12.72	3.05	15.14	2.86	15.10	1.80	13.56	3.27	15.85	3.47	16.44	1.88	14.61
Alternative	ANN	2.78	14.13	2.58	13.83	1.74	12.71	3.15	15.41	2.84	14.82	1.84	13.86	3.22	15.55	3.28	17.02	1.96	15.57
	EP-NN	3.77	18.08	2.69	14.94	1.78	13.12	5.19	20.48	3.53	16.40	1.87	13.90	6.26	20.49	4.58	17.55	1.93	15.29
	Bagging	8.11	70.10	9.03	82.02	7.70	94.31	7.65	64.92	8.36	79.22	7.58	92.71	7.36	68.13	7.79	81.35	7.25	88.69
	Componentwise Boosting	2.69	14.36	2.51	14.32	1.76	13.87	2.97	15.18	2.79	15.57	1.83	14.75	3.21	16.01	3.19	17.89	1.98	16.45
	AdaBoost	3.41	20.62	3.53	22.67	3.14	29.29	3.40	20.48	3.51	23.25	3.22	29.96	3.38	19.76	3.49	22.52	3.17	29.27
APM	c-APM (Constant)	2.65	13.33	2.45	13.05	1.75	12.62	3.13	15.32	2.84	15.00	1.89	14.23	3.78	18.65	3.60	19.09	2.30	19.78
	c-APM (Q-learning)	2.66	13.40	2.45	13.23	1.75	12.79	3.08	15.20	2.84	14.89	1.89	14.41	3.56	16.93	3.46	17.59	2.20	17.85
	Market for Kernels	2.65	12.64	2.43	12.88	1.75	12.77	3.07	14.93	2.79	15.10	1.87	13.99	3.71	19.57	3.42	17.39	2.04	15.70
	Best Individual	2.60	13.20	2.41	12.83	1.72	12.34	3.01	14.90	2.78	14.48	1.84	13.62	3.70	17.64	3.58	17.45	2.00	15.57
	Median Individual	2.66	13.82	2.46	13.55	1.77	13.12	3.17	17.13	2.87	16.86	1.95	16.28	4.23	26.34	3.85	26.06	2.57	25.71
	Worst Individual	5.39	41.21	5.02	41.08	3.94	42.54	5.39	41.37	5.00	41.25	3.96	42.78	5.41	42.11	5.05	42.03	4.01	43.75

Note 1: The RMSE measure is scaled up by the order of 10^4.

Note 2: The MAE measure delivers similar message as the RMSE and so we omitted it from the table for space reasons.

Table 5.7: P-values from the DM test of equal forecast accuracy: equal weights against the remaining combinations of forecasts of U.S. Treasury futures RVOL

Class	Forecast Combination Method	TU (2 Year)									FV (5 Year)								
		h = 1			h = 5			h = 22			h = 1			h = 5			h = 22		
		w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500
Simple	Equal Weights	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Bates-Granger (1)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (2)	0.13	0.01^{**}	0.00^{***}	0.34	0.38	0.00^{***}	0.13	$0.02^{\dagger\dagger}$	0.62	0.06^{*}	0.12	0.00^{***}	0.04^{**}	0.38	0.00^{***}	0.33	$0.03^{\dagger\dagger}$	0.03^{**}
	Bates-Granger (3)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (4)	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^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (5)	0.96	0.38	$0.02^{\dagger\dagger}$	0.13	0.33	0.61	0.81	0.92	0.46	0.05^{**}	0.08^{*}	0.12	0.05^{**}	0.02^{**}	0.00^{***}	0.11	0.22	0.83
	Granger-Ramanathan (1)	$0.08^{\dagger}$	0.51	0.00^{***}	$0.08^{\dagger}$	0.40	0.13	0.26	0.62	0.00^{***}	$0.00^{\dagger\dagger\dagger}$	$0.06^{\dagger}$	0.35	$0.01^{\dagger\dagger\dagger}$	0.20	0.59	0.56	0.71	0.00^{***}
	Granger-Ramanathan (2)	0.11	0.47	0.00^{***}	$0.05^{\dagger}$	0.18	0.01^{***}	0.20	0.55	0.00^{***}	$0.00^{\dagger\dagger\dagger}$	$0.06^{\dagger}$	0.54	$0.06^{\dagger}$	0.11	0.21	0.53	0.57	0.00^{***}
	Granger-Ramanathan (3)	$0.05^{\dagger\dagger}$	0.16	0.00^{***}	$0.09^{\dagger}$	0.29	0.00^{***}	0.29	0.28	0.00^{***}	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger}$	0.92	$0.01^{\dagger\dagger}$	0.14	0.22	0.84	0.81	0.00^{***}
	AFTER	0.01^{**}	0.07^{*}	0.00^{***}	0.12	0.32	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.06^{*}	0.00^{***}	0.00^{***}	0.03^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Median Forecast	0.04^{**}	0.11	0.00^{***}	0.97	1.00	0.23	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.54	0.56	0.45	$0.07^{\dagger}$	0.18	0.74	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$
	Trimmed Mean Forecast	0.25	0.22	0.06^{*}	0.93	0.82	0.17	0.50	0.10	0.27	0.05^{*}	0.04^{**}	0.06^{*}	0.86	0.09^{*}	0.81	0.63	0.67	0.58
	PEW	0.13	0.29	0.00^{***}	0.01^{***}	0.21	0.44	0.00^{***}	0.06^{*}	$0.10^{\dagger}$	0.67	0.76	0.39	0.01^{***}	0.04^{**}	0.77	0.00^{***}	0.00^{***}	0.48
Factor An.	Principal Component Forecast	0.05^{*}	0.21	0.00^{***}	0.10	0.56	0.00^{***}	0.00^{***}	0.23	0.00^{***}	0.13	0.36	0.66	0.02^{**}	0.20	0.25	0.00^{***}	0.00^{***}	0.07^{*}
	Principal Component Forecast (AIC)	0.33	0.49	0.00^{***}	0.25	0.57	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.93	0.78	0.17	0.62	0.37	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}
	Principal Component Forecast (BIC)	0.46	0.37	0.00^{***}	0.27	0.72	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.88	0.71	0.03^{**}	0.66	0.38	0.04^{**}	0.00^{***}	0.00^{***}	0.00^{***}
Shrinkage	Empirical Bayes Estimator	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.01^{\dagger\dagger}$	$0.05^{\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.11	0.13	0.24	$0.04^{\dagger\dagger}$
	Kappa-Shrinkage	0.11	0.60	0.00^{***}	0.14	0.48	0.11	0.11	0.50	0.00^{***}	$0.00^{\dagger\dagger\dagger}$	$0.08^{\dagger}$	0.39	$0.02^{\dagger\dagger}$	0.26	0.53	0.93	0.53	0.00^{***}
	Two-Step Egalitarian LASSO	0.85	0.58	0.42	0.39	0.86	0.26	0.00^{***}	0.08^{*}	0.01^{**}	$0.00^{\dagger\dagger\dagger}$	0.37	0.92	$0.02^{\dagger\dagger}$	0.96	0.79	0.02^{**}	0.02^{**}	0.34
BMA	BMA (Marginal Likelihood)	0.32	0.28	0.00^{***}	0.17	1.00	0.00^{***}	0.00^{***}	0.04^{**}	0.00^{***}	0.51	0.92	0.06^{*}	0.59	0.22	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	BMA (Predictive Likelihood)	0.62	0.48	0.00^{***}	0.26	0.77	0.00^{***}	0.00^{***}	0.44	0.00^{***}	$0.02^{\dagger\dagger}$	0.38	0.03^{**}	0.80	0.46	0.00^{***}	0.00^{***}	0.83	0.00^{***}
Alternative	ANN	0.28	0.10	0.00^{***}	0.84	0.97	0.00^{***}	0.00^{***}	0.20	0.00^{***}	0.19	0.13	0.22	0.37	0.55	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	EP-NN	$0.00^{\dagger\dagger\dagger}$	$0.05^{\dagger\dagger}$	0.12	$0.00^{\dagger\dagger\dagger}$	0.13	0.04^{**}	0.16	0.42	0.00^{***}	0.20	$0.00^{\dagger\dagger\dagger}$	0.19	$0.04^{\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.16	0.10	0.41	0.00^{***}
	Bagging	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$
	Componentwise Boosting	0.58	0.79	0.00^{***}	0.05^{*}	0.64	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	$0.07^{\dagger}$	$0.02^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.14	0.42	0.69	0.00^{***}	0.00^{***}	0.00^{***}
	AdaBoost	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.00^{***}	0.50	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.00^{***}	0.01^{**}	$0.00^{\dagger\dagger\dagger}$
APM	c-APM (Constant)	0.12	0.23	0.00^{***}	0.92	0.82	0.01^{***}	0.01^{***}	0.66	0.92	0.50	0.78	0.21	0.55	0.46	0.23	0.12	0.05^{*}	$0.01^{\dagger\dagger}$
	c-APM (Q-learning)	0.17	0.25	0.00^{***}	0.47	0.81	0.02^{**}	0.00^{***}	0.09^{*}	0.01^{***}	0.53	0.28	0.33	0.42	0.44	0.71	0.02^{**}	0.00^{***}	0.10
	Market for Kernels	0.00^{***}	0.04^{**}	0.00^{***}	0.00^{***}	0.07^{*}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.05^{**}	0.51	0.19	0.00^{***}	0.06^{*}	0.10^{*}	0.00^{***}	0.00^{***}	0.00^{***}

*The forecasts are significantly better than forecasts from the equal weights
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
†The forecasts are significantly worse than forecasts from the equal weights
$^{\dagger\dagger\dagger}p < 0.01$, $^{\dagger\dagger}p < 0.05$, $^{\dagger}p < 0.1$

Table 5.8: P-values from the DM test of equal forecast accuracy: equal weights against the remaining combinations of forecasts of U.S. Treasury futures RVOL

Class	Forecast Combination Method	TY (10 Year)									US (30 Year)								
		h = 1			h = 5			h = 22			h = 1			h = 5			h = 22		
		w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500
Simple	Equal Weights	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Bates-Granger (1)	0.03^{**}	0.25	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.14	0.50	0.01^{***}	0.00^{***}	0.07^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (2)	0.16	0.74	0.12	0.42	0.83	0.42	$0.02^{\dagger\dagger}$	$0.04^{\dagger\dagger}$	0.85	0.61	0.73	0.19	0.05^{**}	0.49	0.28	0.25	0.00^{***}	$0.06^{\dagger}$
	Bates-Granger (3)	0.01^{***}	0.12	0.00^{***}	0.00^{***}	0.01^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.06^{*}	0.32	0.00^{***}	0.00^{***}	0.04^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (4)	0.01^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.04^{**}	0.05^{*}	0.15	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (5)	0.84	0.91	0.13	1.00	0.84	0.73	0.78	0.49	0.44	0.88	0.93	0.50	0.27	0.21	0.04^{**}	0.83	0.89	0.85
	Granger-Ramanathan (1)	$0.00^{\dagger\dagger\dagger}$	$0.04^{\dagger\dagger}$	$0.03^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.13	0.18	0.72	0.70	0.00^{***}	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.58	$0.03^{\dagger\dagger}$	0.17	0.68	0.83	0.17	0.03^{**}
	Granger-Ramanathan (2)	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.33	$0.00^{\dagger\dagger\dagger}$	0.29	0.31	0.93	0.13	0.00^{***}	$0.00^{\dagger\dagger\dagger}$	$0.02^{\dagger\dagger}$	0.94	$0.00^{\dagger\dagger\dagger}$	0.22	0.74	0.71	0.31	0.03^{**}
	Granger-Ramanathan (3)	$0.02^{\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	$0.09^{\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.07^{\dagger}$	0.12	0.06^{*}	0.94	0.00^{***}	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.66	$0.05^{\dagger}$	$0.09^{\dagger}$	0.55	0.91	0.21	0.01^{***}
	AFTER	0.28	0.69	0.00^{***}	0.24	0.52	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.94	0.87	0.00^{***}	0.54	0.71	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}
	Median Forecast	0.54	0.62	0.31	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	$0.04^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.08^{\dagger}$	$0.09^{\dagger}$	$0.02^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$
	Trimmed Mean Forecast	0.56	0.62	0.58	0.22	0.41	0.64	0.36	0.29	0.97	0.37	0.36	0.84	0.08^{*}	0.10	0.06^{*}	0.65	0.26	0.24
	PEW	0.93	0.66	0.40	0.11	0.24	0.15	0.00^{***}	0.00^{***}	$0.09^{\dagger}$	0.86	0.48	0.51	0.18	0.57	0.19	0.00^{***}	0.00^{***}	$0.05^{\dagger}$
Factor An.	Principal Component Forecast	0.22	0.62	0.49	0.04^{**}	0.36	0.20	0.00^{***}	0.02^{**}	0.26	0.29	0.49	0.80	0.04^{**}	0.21	0.61	0.00^{***}	0.01^{***}	0.52
	Principal Component Forecast (AIC)	0.47	0.61	0.83	0.85	0.32	0.70	0.00^{***}	0.00^{***}	0.43	0.92	0.95	0.42	0.65	0.26	0.76	0.00^{***}	0.00^{***}	0.37
	Principal Component Forecast (BIC)	0.69	0.69	0.86	0.93	0.40	0.86	0.00^{***}	0.00^{***}	0.33	0.96	0.74	0.52	0.32	0.59	0.76	0.00^{***}	0.00^{***}	0.34
Shrinkage	Empirical Bayes Estimator	$0.03^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.10	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.25
	Kappa-Shrinkage	$0.00^{\dagger\dagger\dagger}$	$0.06^{\dagger}$	$0.03^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.21	0.20	0.30	0.53	0.00^{***}	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.63	$0.05^{\dagger\dagger}$	0.25	0.74	0.37	0.09^{*}	0.02^{**}
	Two-Step Egalitarian LASSO	$0.00^{\dagger\dagger\dagger}$	0.34	0.63	$0.04^{\dagger\dagger}$	0.93	0.51	0.05^{**}	0.01^{**}	0.97	$0.07^{\dagger}$	0.13	$0.02^{\dagger\dagger}$	0.83	0.60	0.34	0.00^{***}	0.00^{***}	0.00^{***}
BMA	BMA (Marginal Likelihood)	0.40	0.77	0.80	0.52	0.23	0.46	0.00^{***}	0.00^{***}	0.00^{***}	0.43	0.25	0.67	0.42	0.46	0.96	0.00^{***}	0.00^{***}	0.00^{***}
	BMA (Predictive Likelihood)	$0.04^{\dagger\dagger}$	0.42	0.27	0.30	0.70	0.03^{**}	0.00^{***}	0.37	0.00^{***}	$0.07^{\dagger}$	0.61	0.38	$0.08^{\dagger}$	0.77	0.39	0.18	0.78	0.00^{***}
Alternative	ANN	$0.04^{\dagger\dagger}$	0.55	0.41	0.35	0.20	0.11	0.02^{**}	0.00^{***}	0.00^{***}	0.13	0.97	$0.04^{\dagger\dagger}$	0.20	0.72	0.77	0.26	0.00^{***}	0.02^{**}
	EP-NN	$0.10^{\dagger}$	$0.01^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.27	0.18	0.17	$0.05^{\dagger}$	0.33	0.00^{***}	0.15	$0.07^{\dagger}$	$0.03^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.03^{\dagger\dagger}$	0.15	0.02^{**}
	Bagging	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$
	Componentwise Boosting	$0.04^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.50	0.83	$0.01^{\dagger\dagger\dagger}$	0.00^{***}	0.00^{***}	0.02^{**}	$0.01^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.61	0.59	$0.00^{\dagger\dagger\dagger}$	0.00^{***}	0.00^{***}	0.05^{*}
	AdaBoost	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.00^{***}	0.07^{*}	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.00^{***}	0.99	$0.00^{\dagger\dagger\dagger}$
APM	c-APM (Constant)	0.86	0.33	$0.04^{\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.45	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.10^{\dagger}$	$0.09^{\dagger}$	$0.01^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$
	c-APM (Q-learning)	0.15	0.31	$0.02^{\dagger\dagger}$	0.94	$0.07^{\dagger}$	0.49	0.00^{***}	0.15	$0.01^{\dagger\dagger\dagger}$	$0.05^{\dagger}$	$0.10^{\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.12	0.12	$0.00^{\dagger\dagger\dagger}$	0.55	0.61	0.71
	Market for Kernels	0.32	0.96	0.20	0.00^{***}	0.12	0.59	0.00^{***}	0.09^{*}	0.66	0.73	0.45	0.74	0.04^{**}	0.25	0.77	0.04^{**}	0.09^{*}	0.87

*The forecasts are significantly better than forecasts from the equal weights
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
†The forecasts are significantly worse than forecasts from the equal weights
$^{\dagger\dagger\dagger}p < 0.01$, $^{\dagger\dagger}p < 0.05$, $^{\dagger}p < 0.1$

Table 5.9: P-values from the DM test of equal forecast accuracy: Market for Kernels against the remaining combinations of forecasts of U.S. Treasury futures RVOL

Class	Forecast Combination Method	TU (2 Year)									FV (5 Year)								
		h = 1			h = 5			h = 22			h = 1			h = 5			h = 22		
		w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500
Simple	Equal Weights	0.00^{***}	0.04^{**}	0.00^{***}	0.00^{***}	0.07^{*}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.05^{**}	0.51	0.19	0.00^{***}	0.06^{*}	0.10^{*}	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (1)	0.74	0.54	0.28	0.29	0.59	0.60	0.01^{***}	0.31	0.01^{**}	0.88	0.69	0.55	0.44	0.66	0.78	0.00^{***}	0.03^{**}	0.00^{***}
	Bates-Granger (2)	0.07^{*}	0.50	0.74	0.00^{***}	0.27	0.59	0.00^{***}	0.01^{***}	0.00^{***}	0.62	0.62	0.58	0.01^{**}	0.06^{*}	0.81	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (3)	0.01^{***}	0.18	0.02^{**}	0.03^{**}	0.26	0.02^{**}	0.00^{***}	0.07^{*}	0.00^{***}	0.19	0.80	0.62	0.01^{***}	0.18	0.27	0.00^{***}	0.00^{***}	0.00^{***}
	Bates-Granger (4)	$0.05^{\dagger\dagger}$	0.30	$0.03^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.03^{\dagger\dagger}$	$0.02^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.24	$0.01^{\dagger\dagger\dagger}$	$0.07^{\dagger}$	$0.10^{\dagger}$	$0.03^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger}$	0.24	$0.00^{\dagger\dagger\dagger}$
	Bates-Granger (5)	0.00^{***}	0.01^{**}	0.00^{***}	0.00^{***}	0.01^{**}	0.00^{***}	0.00^{***}	0.01^{**}	0.00^{***}	0.85	0.72	0.58	0.01^{**}	0.44	0.63	0.00^{***}	0.00^{***}	0.00^{***}
	Granger-Ramanathan (1)	0.02^{**}	0.03^{**}	0.76	0.03^{**}	0.26	0.15	0.69	0.85	0.47	0.00^{***}	0.01^{***}	0.06^{*}	0.00^{***}	0.08^{*}	0.66	0.15	0.41	0.30
	Granger-Ramanathan (2)	0.04^{**}	0.06^{*}	0.16	0.03^{**}	0.05^{*}	0.88	0.08^{*}	0.16	$0.00^{\dagger\dagger\dagger}$	0.00^{***}	0.01^{***}	0.68	0.02^{**}	0.02^{**}	0.75	0.15	0.60	0.12
	Granger-Ramanathan (3)	0.01^{**}	0.01^{***}	$0.08^{\dagger}$	0.05^{**}	0.19	$0.03^{\dagger\dagger}$	0.74	0.64	$0.03^{\dagger\dagger}$	0.00^{***}	0.00^{***}	0.34	0.00^{***}	0.06^{*}	0.73	0.30	0.57	$0.02^{\dagger\dagger}$
	AFTER	0.00^{***}	0.04^{**}	0.00^{***}	0.00^{***}	0.07^{*}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.05^{**}	0.51	0.19	0.00^{***}	0.06^{*}	0.10^{*}	0.00^{***}	0.00^{***}	0.00^{***}
	Median Forecast	0.67	0.77	0.00^{***}	0.01^{***}	0.01^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.03^{**}	0.67	0.41	0.00^{***}	0.00^{***}	0.08^{*}	0.00^{***}	0.00^{***}	0.00^{***}
	Trimmed Mean Forecast	0.00^{***}	0.04^{**}	0.00^{***}	0.00^{***}	0.07^{*}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.05^{**}	0.51	0.19	0.00^{***}	0.06^{*}	0.10^{*}	0.00^{***}	0.00^{***}	0.00^{***}
	PEW	0.69	0.52	0.45	$0.02^{\dagger\dagger}$	0.59	0.04^{**}	$0.03^{\dagger\dagger}$	0.31	0.00^{***}	0.74	0.76	0.82	$0.10^{\dagger}$	0.27	0.13	$0.00^{\dagger\dagger\dagger}$	$0.03^{\dagger\dagger}$	0.00^{***}
Factor An.	Principal Component Forecast	0.35	0.91	0.59	0.23	0.92	0.54	$0.08^{\dagger}$	0.82	0.03^{**}	0.58	0.66	0.49	0.11	0.52	0.84	$0.02^{\dagger\dagger}$	0.29	0.09^{*}
	Principal Component Forecast (AIC)	0.98	0.39	0.34	0.52	0.98	0.21	$0.02^{\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.78	0.53	0.28	0.78	0.88	0.76	0.14	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.70
	Principal Component Forecast (BIC)	0.75	0.61	0.63	0.53	0.82	0.21	$0.01^{\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.73	0.47	0.95	0.46	0.82	0.90	0.43	$0.01^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.53
Shrinkage	Empirical Bayes Estimator	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.02^{**}	0.00^{***}	0.05^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.10^{*}	0.13	0.24	0.02^{**}
	Kappa-Shrinkage	0.03^{**}	0.04^{**}	0.69	0.07^{*}	0.31	0.18	0.40	0.99	0.51	0.00^{***}	0.01^{**}	0.06^{*}	0.01^{***}	0.11	0.71	0.30	0.53	0.26
	Two-Step Egalitarian LASSO	0.06^{*}	0.00^{***}	0.00^{***}	0.95	0.60	0.03^{**}	$0.02^{\dagger\dagger}$	0.41	0.00^{***}	0.00^{***}	0.13	0.45	0.01^{***}	0.48	0.34	0.66	0.77	0.03^{**}
BMA	BMA (Marginal Likelihood)	0.93	0.56	$0.00^{\dagger\dagger\dagger}$	0.39	0.67	$0.02^{\dagger\dagger}$	$0.02^{\dagger\dagger}$	0.16	$0.03^{\dagger\dagger}$	0.19	0.51	0.34	0.92	0.52	$0.01^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	$0.03^{\dagger\dagger}$	$0.01^{\dagger\dagger}$
	BMA (Predictive Likelihood)	0.06^{*}	0.30	0.46	0.75	0.44	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger}$	0.79	$0.00^{\dagger\dagger\dagger}$	0.00^{***}	0.27	0.18	0.31	1.00	$0.00^{\dagger\dagger\dagger}$	0.28	0.25	$0.00^{\dagger\dagger\dagger}$
Alternative	ANN	0.01^{**}	0.02^{**}	0.56	0.44	0.68	0.17	$0.02^{\dagger\dagger}$	0.55	$0.09^{\dagger}$	0.04^{**}	0.03^{**}	0.64	0.11	0.97	$0.04^{\dagger\dagger}$	0.15	$0.01^{\dagger\dagger\dagger}$	0.25
	EP-NN	0.00^{***}	0.00^{***}	0.09^{*}	0.00^{***}	0.10	1.00	0.14	0.33	$0.02^{\dagger\dagger}$	0.20	0.00^{***}	0.15	0.04^{**}	0.01^{***}	0.61	0.09^{*}	0.19	$0.02^{\dagger\dagger}$
	Bagging	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^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Componentwise Boosting	0.20	0.15	0.51	0.20	0.93	$0.08^{\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.14	0.25	0.02^{**}	0.03^{**}	0.00^{***}	0.60	0.98	0.16	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger}$	0.35
	AdaBoost	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	$0.06^{\dagger}$	0.72	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	$0.01^{\dagger\dagger}$	0.65	0.00^{***}
APM	c-APM (Constant)	0.89	0.20	0.02^{**}	0.04^{**}	0.06^{*}	0.05^{*}	0.06^{*}	0.02^{**}	0.00^{***}	0.24	0.00^{***}	0.01^{**}	0.00^{***}	0.01^{**}	0.01^{**}	0.00^{***}	0.01^{***}	0.00^{***}
	c-APM (Q-learning)	0.74	0.26	0.07^{*}	0.76	0.06^{*}	0.04^{**}	$0.05^{\dagger}$	0.54	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{***}	0.18	0.00^{***}	0.21	0.02^{**}	0.09^{*}	0.00^{***}
	Market for Kernels	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

*The Market for Kernels forecasts are significantly better than the given forecasts

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

†The Market for Kernels forecasts are significantly worse than the given forecasts

$^{\dagger\dagger\dagger}p < 0.01$, $^{\dagger\dagger}p < 0.05$, $^{\dagger}p < 0.1$

Table 5.10: P-values from the DM test of equal forecast accuracy: Market for Kernels against the remaining combinations of forecasts of U.S. Treasury futures RVOL

Class	Forecast Combination Method	TY (10 Year)									US (30 Year)								
		h = 1			h = 5			h = 22			h = 1			h = 5			h = 22		
		w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500	w=100	w=200	w=500
Simple	Equal Weights	0.32	0.96	0.20	0.00^{***}	0.12	0.59	0.00^{***}	0.09^{*}	0.66	0.73	0.45	0.74	0.04^{**}	0.25	0.77	0.04^{**}	0.09^{*}	0.87
	Bates-Granger (1)	0.49	0.44	0.86	0.13	0.31	0.87	0.00^{***}	0.55	0.98	0.20	0.17	0.57	0.45	0.42	0.55	0.40	0.25	0.74
	Bates-Granger (2)	0.60	0.55	0.64	0.01^{***}	0.07^{*}	0.39	0.00^{***}	0.06^{*}	0.68	0.36	0.19	0.52	0.87	0.13	0.58	0.03^{**}	0.14	0.94
	Bates-Granger (3)	0.67	0.73	0.42	0.02^{**}	0.18	0.80	0.00^{***}	0.20	0.78	0.47	0.31	0.96	0.11	0.31	0.68	0.11	0.14	0.82
	Bates-Granger (4)	0.14	0.16	0.49	$0.01^{\dagger\dagger}$	0.10	$0.00^{\dagger\dagger\dagger}$	0.36	$0.09^{\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.23	0.15	0.61	$0.01^{\dagger\dagger\dagger}$	0.13	$0.01^{\dagger\dagger}$	$0.03^{\dagger\dagger}$	0.17	$0.00^{\dagger\dagger\dagger}$
	Bates-Granger (5)	0.44	1.00	0.06^{*}	0.01^{***}	0.15	0.67	0.00^{***}	0.10	0.73	0.76	0.54	0.52	0.35	0.54	0.49	0.04^{**}	0.09^{*}	0.89
	Granger-Ramanathan (1)	0.00^{***}	0.08^{*}	0.00^{***}	0.00^{***}	0.05^{**}	0.08^{*}	0.45	0.84	$0.01^{\dagger\dagger\dagger}$	0.00^{***}	0.00^{***}	0.51	0.02^{**}	0.07^{*}	0.83	0.60	0.55	$0.02^{\dagger\dagger}$
	Granger-Ramanathan (2)	0.00^{***}	0.00^{***}	0.07^{*}	0.00^{***}	0.13	0.16	0.25	0.42	$0.01^{\dagger\dagger\dagger}$	0.00^{***}	0.01^{***}	0.94	0.00^{***}	0.09^{*}	0.55	0.72	0.78	$0.01^{\dagger\dagger}$
	Granger-Ramanathan (3)	0.01^{**}	0.03^{**}	0.01^{**}	0.00^{***}	0.02^{**}	0.05^{*}	0.78	0.78	$0.00^{\dagger\dagger\dagger}$	0.00^{***}	0.00^{***}	0.56	0.04^{**}	0.03^{**}	0.71	0.59	0.62	$0.00^{\dagger\dagger\dagger}$
	AFTER	0.32	0.96	0.20	0.00^{***}	0.12	0.59	0.00^{***}	0.09^{*}	0.66	0.73	0.45	0.74	0.04^{**}	0.25	0.77	0.04^{**}	0.09^{*}	0.87
	Median Forecast	0.01^{***}	0.24	0.01^{***}	0.00^{***}	0.00^{***}	0.09^{*}	0.00^{***}	0.00^{***}	0.10	0.00^{***}	0.01^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.25	0.00^{***}	0.00^{***}	0.15
	Trimmed Mean Forecast	0.32	0.96	0.20	0.00^{***}	0.12	0.59	0.00^{***}	0.09^{*}	0.66	0.73	0.45	0.74	0.04^{**}	0.25	0.77	0.04^{**}	0.09^{*}	0.87
	PEW	0.84	0.75	0.11	0.41	0.85	0.11	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.07^{*}	0.81	0.99	0.41	0.38	0.81	0.41	$0.00^{\dagger\dagger\dagger}$	0.19	0.17
Factor An.	Principal Component Forecast	0.36	0.46	0.73	0.16	0.87	0.50	$0.04^{\dagger\dagger}$	$0.05^{\dagger\dagger}$	0.57	0.15	0.11	0.96	$0.09^{\dagger}$	0.46	0.54	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.77
	Principal Component Forecast (AIC)	0.32	0.64	0.24	0.39	0.91	0.97	$0.03^{\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.68	0.99	0.64	0.58	0.93	0.53	0.99	$0.00^{\dagger\dagger\dagger}$	$0.02^{\dagger\dagger}$	0.27
	Principal Component Forecast (BIC)	0.41	0.64	0.44	0.44	0.98	0.55	$0.06^{\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.54	0.94	0.66	0.72	0.56	0.84	0.48	$0.00^{\dagger\dagger\dagger}$	$0.02^{\dagger\dagger}$	0.25
Shrinkage	Empirical Bayes Estimator	0.03^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.10^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{**}	0.00^{***}	0.28
	Kappa-Shrinkage	0.00^{***}	0.12	0.00^{***}	0.00^{***}	0.08^{*}	0.09^{*}	0.84	0.99	$0.01^{\dagger\dagger\dagger}$	0.00^{***}	0.01^{***}	0.55	0.03^{**}	0.12	0.90	0.91	0.39	$0.01^{\dagger\dagger}$
	Two-Step Egalitarian LASSO	0.00^{***}	0.42	0.27	0.00^{***}	0.50	0.87	0.30	$0.08^{\dagger}$	0.75	0.13	0.39	0.02^{**}	0.58	0.88	0.60	$0.00^{\dagger\dagger\dagger}$	$0.08^{\dagger}$	$0.01^{\dagger\dagger\dagger}$
BMA	BMA (Marginal Likelihood)	0.31	0.82	0.23	0.96	0.71	0.70	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.53	0.58	0.85	0.62	0.82	0.71	$0.00^{\dagger\dagger\dagger}$	$0.07^{\dagger}$	$0.00^{\dagger\dagger\dagger}$
	BMA (Predictive Likelihood)	0.04^{**}	0.55	0.73	0.08^{*}	0.30	0.10	0.14	0.84	$0.00^{\dagger\dagger\dagger}$	0.13	0.88	0.57	0.05^{*}	0.44	0.24	0.62	0.61	$0.00^{\dagger\dagger\dagger}$
Alternative	ANN	0.03^{**}	0.66	0.17	0.30	0.68	0.23	0.46	$0.00^{\dagger\dagger\dagger}$	$0.01^{\dagger\dagger\dagger}$	0.21	0.84	0.04^{**}	0.18	0.58	0.56	0.13	$0.02^{\dagger\dagger}$	$0.02^{\dagger\dagger}$
	EP-NN	0.10^{*}	0.01^{**}	0.00^{***}	0.27	0.16	0.10^{*}	0.04^{**}	0.30	$0.00^{\dagger\dagger\dagger}$	0.15	0.07^{*}	0.03^{**}	0.00^{***}	0.00^{***}	0.03^{**}	0.02^{**}	0.12	$0.01^{\dagger\dagger}$
	Bagging	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^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	Componentwise Boosting	0.05^{**}	0.04^{**}	0.00^{***}	0.91	0.60	0.02^{**}	$0.00^{\dagger\dagger\dagger}$	$0.00^{\dagger\dagger\dagger}$	0.26	0.05^{**}	0.03^{**}	0.00^{***}	0.94	0.36	0.01^{***}	$0.00^{\dagger\dagger\dagger}$	0.20	0.12
	AdaBoost	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	$0.03^{\dagger\dagger}$	0.59	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	$0.05^{\dagger\dagger}$	0.17	0.00^{***}
APM	c-APM (Constant)	0.47	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{**}	0.00^{***}	0.01^{**}	0.19	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.34	0.00^{***}	0.01^{***}	0.18
	c-APM (Q-learning)	0.01^{***}	0.02^{**}	0.00^{***}	0.08^{*}	0.00^{***}	0.12	0.62	0.66	0.00^{***}	0.01^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.11	0.03^{**}	0.81
	Market for Kernels	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

*The Market for Kernels forecasts are significantly better than the given forecasts
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
†The Market for Kernels forecasts are significantly worse than the given forecasts
$^{\dagger\dagger\dagger}p < 0.01$, $^{\dagger\dagger}p < 0.05$, $^{\dagger}p < 0.1$

of the Market for Kernels on very low significance levels in favour of the Market for Kernels vast majority of the method on almost all the datasets. The only exception is the HAR method, the best individual forecast of the realized volatility, for which the we cannot reject the hypothesis of equal forecast accuracy on all datasets excluding the TU with 1-day ahead horizons for all the rolling window lengths. This simply indicates that in most cases, the Market for Kernels learns to put the heaviest weight on the HAR model and thus their forecasts converge.

5.5 Forecast Combination Ranking

In this section we make an attempt for a fusion of results from both applications and a simple assessment of overall out-of-sample forecasting performance of the forecast combination methods across all the studied datasets. We present the table of average rankings 5.12, where for each dataset, the ranks are assigned to the combining methods according to the individual accuracy measures RMSE, MAE and MAPE and then averaged into composite ranks. These ranks are further averaged across all the rolling window lengths for all the given datasets and then presented in the table. The subtotals then represent averages of all these ranks across all the datasets in each given application. The total rank is then a simple average of these subtotals rather than the individual ranks as we do want to put disproportionately more weight on the financial application, which covers more datasets.

The table signifies the results that were already presented in the previous sections. Within the ECB SPF application, the best performing methods are the simple Bates-Granger optimal-combining weights, the AFTER method and the artificial prediction markets. The factor analytic show a solid performance in the ECB application and the best performance as a class in the realized volatility application. The only exception make the cases of large samples and long horizons, where the BMA forecast combination methods perform better. The Market for Kernels method shows very decent performance in both applications and is the second overall best combination method. The overall best combination method, which dominates all the other methods in both applications is the Bates-Granger (4). The Empirical Bayes Estimator method and the alternative methods, excluding the simple ANN, are among the methods with the worst overall performance.

Table 5.11: P-values from the DM test of equal forecast accuracy: Market for Kernels against the individual forecasts of U.S. Treasury futures RVOL

Future	Volatility Model	h = 1				h = 5				h = 22			
		TU	FV	TY	US	TU	FV	TY	US	TU	FV	TY	US
w = 100	Historical Volatility	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	RiskMetrics	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.05^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.07^{*}	0.01^{**}	0.01^{***}
	HAR	$0.00^{\dagger\dagger\dagger}$	$0.05^{\dagger}$	0.54	0.27	$0.08^{\dagger}$	0.20	0.68	0.53	0.40	0.39	0.15	0.10
	GARCH	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.02^{**}	0.01^{**}	0.00^{***}	0.00^{***}	0.03^{**}	0.09^{*}
	VAR (TU)	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.07^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{***}
	VAR (TU, FV)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.10	0.00^{***}	0.00^{***}	0.05^{**}	0.02^{**}
	VAR (TU, TY)	0.00^{***}	0.01^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{***}	0.15	0.00^{***}	0.00^{***}	0.01^{**}	0.08^{*}
	VAR (TU, US)	0.00^{***}	0.06^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.05^{**}	0.14	0.00^{***}	0.00^{***}	0.34	0.22
	VAR (TU, FV, TY)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.01^{***}	0.03^{**}
	VAR (TU, FV, US)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{**}	0.01^{***}	0.00^{***}	0.00^{***}	0.22	0.21
	VAR (TU, TY, US)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.21	0.28
	VAR (TU, FV, TY, US)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.04^{**}	0.05^{*}
w = 200	Historical Volatility	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	RiskMetrics	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.05^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.07^{*}	0.01^{**}	0.01^{***}
	HAR	$0.00^{\dagger\dagger\dagger}$	$0.05^{\dagger}$	0.54	0.27	$0.08^{\dagger}$	0.20	0.68	0.53	0.40	0.39	0.15	0.10
	GARCH	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.02^{**}	0.01^{**}	0.00^{***}	0.00^{***}	0.03^{**}	0.09^{*}
	VAR (TU)	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.07^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{***}
	VAR (TU, FV)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.10	0.00^{***}	0.00^{***}	0.05^{**}	0.02^{**}
	VAR (TU, TY)	0.00^{***}	0.01^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{***}	0.15	0.00^{***}	0.00^{***}	0.01^{**}	0.08^{*}
	VAR (TU, US)	0.00^{***}	0.06^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.05^{**}	0.14	0.00^{***}	0.00^{***}	0.34	0.22
	VAR (TU, FV, TY)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.01^{***}	0.03^{**}
	VAR (TU, FV, US)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{**}	0.01^{***}	0.00^{***}	0.00^{***}	0.22	0.21
	VAR (TU, TY, US)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.21	0.28
	VAR (TU, FV, TY, US)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.04^{**}	0.05^{*}
w = 500	Historical Volatility	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	RiskMetrics	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.05^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.07^{*}	0.01^{**}	0.01^{***}
	HAR	$0.00^{\dagger\dagger\dagger}$	$0.05^{\dagger}$	0.54	0.27	$0.08^{\dagger}$	0.20	0.68	0.53	0.40	0.39	0.15	0.10
	GARCH	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.02^{**}	0.01^{**}	0.00^{***}	0.00^{***}	0.03^{**}	0.09^{*}
	VAR (TU)	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.07^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{***}
	VAR (TU, FV)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.10	0.00^{***}	0.00^{***}	0.05^{**}	0.02^{**}
	VAR (TU, TY)	0.00^{***}	0.01^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{***}	0.15	0.00^{***}	0.00^{***}	0.01^{**}	0.08^{*}
	VAR (TU, US)	0.00^{***}	0.06^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.05^{**}	0.14	0.00^{***}	0.00^{***}	0.34	0.22
	VAR (TU, FV, TY)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.01^{***}	0.03^{**}
	VAR (TU, FV, US)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.01^{**}	0.01^{***}	0.00^{***}	0.00^{***}	0.22	0.21
	VAR (TU, TY, US)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.02^{**}	0.00^{***}	0.00^{***}	0.21	0.28
	VAR (TU, FV, TY, US)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.04^{**}	0.05^{*}

*The Market for Kernels forecasts are significantly better than the given forecasts

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

†The Market for Kernels forecasts are significantly worse than the given forecasts

$^{\dagger\dagger\dagger}p < 0.01$, $^{\dagger\dagger}p < 0.05$, $^{\dagger}p < 0.1$

Table 5.12: Average ranks of forecast combinations methods across all the datasets obtained by averaging the ranks based on RMSE, MAE and MAPE

Class	Forecast Combination Method	Total	ECB SPF							U.S. Treasury Futures RVOL												
			Subtotal	RGDP		HICP		UNEM		Subtotal	h = 1				h = 5				h = 22			
				1Y	2Y	1Y	2Y	1Y	2Y		TU	FV	TY	US	TU	FV	TY	US	TU	FV	TY	US
Simple	Equal Weights	15.6	12.9	13.9	15.5	14.4	16.8	5.3	11.4	18.2	11.1	19.2	15.1	13.5	14.9	21.6	17.7	16.1	22.6	22.5	23.1	21.7
	Bates-Granger (1)	11.4	10.9	10.4	12.4	11.1	14.9	7.4	8.9	12.0	5.4	9.4	7.8	7.4	11.2	13.0	11.2	10.7	16.8	17.2	17.2	16.8
	Bates-Granger (2)	11.1	7.5	7.6	10.1	5.2	11.0	5.4	5.7	14.7	4.4	10.0	7.8	8.3	6.9	16.0	16.7	13.1	24.8	22.9	23.7	21.8
	Bates-Granger (3)	13.0	11.9	12.0	13.7	12.4	16.0	6.8	10.3	14.2	6.3	14.7	11.1	10.2	9.4	16.9	13.2	12.6	19.1	19.1	18.8	18.6
	Bates-Granger (4)	5.9	5.9	4.7	9.6	7.1	9.3	1.4	3.1	5.9	3.6	4.9	4.8	5.8	2.7	4.8	3.9	2.7	6.1	10.8	11.2	9.8
	Bates-Granger (5)	13.0	8.1	7.1	11.7	5.8	10.1	4.2	9.9	17.9	15.4	13.6	17.6	15.1	17.4	17.0	16.0	12.2	23.7	23.8	20.2	22.4
	Granger-Ramanathan (1)	20.2	23.6	25.3	26.8	19.6	20.3	23.8	25.7	16.8	20.0	20.4	21.1	19.8	21.9	16.1	19.8	19.0	13.7	11.9	9.4	8.6
	Granger-Ramanathan (2)	18.1	23.1	25.3	26.6	20.6	18.9	22.1	25.2	13.1	15.9	13.6	16.4	15.6	17.4	12.3	13.3	14.4	12.4	9.4	8.6	7.7
	Granger-Ramanathan (3)	19.8	24.5	27.1	27.4	22.3	21.6	23.2	25.2	15.0	17.7	19.6	21.1	18.9	16.3	15.7	21.2	19.8	8.0	8.0	6.9	7.4
	AFTER	12.6	8.5	8.3	7.6	9.2	12.8	8.4	4.4	16.8	9.7	17.7	13.7	12.8	13.4	20.1	16.2	14.6	21.0	21.0	21.6	20.1
	Median Forecast	16.8	13.2	9.0	16.3	13.7	19.3	10.1	10.8	20.4	11.6	13.2	14.1	14.8	21.3	21.3	21.6	21.7	26.2	26.3	26.3	26.2
	Trimmed Mean Forecast	15.4	12.6	11.8	15.5	14.3	16.3	6.4	11.2	18.3	11.2	19.1	15.2	13.7	14.9	21.6	17.8	16.1	22.4	22.5	23.1	21.6
	PEW	13.1	15.4	14.6	10.1	20.1	19.9	14.6	13.2	10.8	10.8	7.3	10.8	12.3	11.2	8.2	8.8	11.7	12.7	10.3	11.4	13.9
Factor An.	Principal Component Forecast	10.1	13.9	13.4	7.8	17.0	12.9	14.7	17.4	6.4	9.7	4.8	2.3	3.1	6.4	4.9	2.0	2.3	10.6	10.1	9.7	11.1
	Principal Component Forecast (AIC)	10.0	12.8	5.8	8.8	19.6	11.8	14.1	17.0	7.1	14.6	6.2	7.0	4.8	8.8	5.2	6.9	8.6	4.2	5.1	6.8	7.6
	Principal Component Forecast (BIC)	10.0	13.7	7.6	8.8	18.4	14.4	16.0	17.2	6.4	13.6	3.7	5.1	4.6	8.0	4.8	7.0	5.2	3.7	5.1	7.6	8.0
Shrinkage	Empirical Bayes Estimator	24.3	21.4	24.4	24.0	20.0	16.6	21.0	22.3	27.1	27.7	27.7	27.7	26.6	27.6	27.9	27.4	27.7	27.9	26.3	27.2	24.2
	Kappa-Shrinkage	13.9	12.7	11.7	19.7	8.9	9.0	15.5	11.4	15.0	18.6	19.1	19.9	18.6	20.3	14.9	18.1	17.6	11.1	10.6	6.6	5.0
	Two-Step Egalitarian LASSO	15.9	14.8	16.8	10.7	16.1	17.9	13.7	13.8	16.9	22.6	19.0	18.3	20.2	19.4	20.6	14.6	14.7	13.7	15.1	14.7	10.0
BMA	BMA (Marginal Likelihood)	11.7	15.1	18.8	6.0	19.1	13.4	16.8	16.8	8.3	18.2	11.1	10.3	12.4	13.2	7.6	5.7	7.7	4.9	2.2	2.4	3.4
	BMA (Predictive Likelihood)	18.0	25.6	25.8	24.6	27.0	22.3	26.4	27.6	10.4	15.0	12.1	14.1	12.8	9.9	7.9	12.3	13.7	5.3	8.2	6.3	6.8
Alternative	ANN	12.7	14.6	21.4	13.1	16.3	14.1	18.8	4.0	10.8	18.2	12.0	14.0	16.6	12.7	8.2	7.8	12.9	5.8	5.9	5.4	9.7
	EP-NN	24.4	27.8	28.7	29.0	28.3	28.6	26.8	25.3	20.9	24.8	24.7	25.1	26.8	20.2	19.7	24.0	24.8	15.2	14.7	15.6	15.7
	Bagging	24.8	20.7	24.7	17.0	13.3	11.3	29.0	29.0	28.9	29.0	29.0	28.9	28.9	28.9	28.9	28.9	29.0	28.9	28.7	28.8	28.9
	Componentwise Boosting	17.1	17.9	17.1	7.4	20.9	12.9	27.0	21.8	16.4	20.7	23.4	23.9	24.1	13.0	16.1	18.2	20.3	8.7	9.6	9.2	9.9
	AdaBoost	20.4	16.4	18.4	20.8	10.9	1.3	25.2	21.7	24.5	27.2	27.0	26.9	27.1	26.9	26.2	26.3	26.3	21.6	17.9	18.6	21.9
APM	c-APM (Constant)	13.7	9.6	6.6	13.3	10.7	15.3	7.3	4.6	17.8	10.4	13.1	13.1	15.4	16.2	16.3	20.2	20.6	18.3	20.6	24.1	25.2
	c-APM (Q-learning)	12.9	11.3	10.6	15.3	4.9	18.1	9.4	9.6	14.5	13.8	12.9	16.8	17.4	15.3	11.2	11.6	13.0	13.2	15.7	16.1	16.9
	Market for Kernels	9.1	8.6	6.1	5.6	7.8	7.8	14.0	10.6	9.5	8.0	6.6	5.0	7.4	8.9	10.0	6.7	6.3	12.6	13.6	14.6	14.3

Chapter 6

Discussion

This chapter is divided into two sections. Firstly, we discuss the results of both the macroeconomic and the financial application results in the context of the literature of other researches. Then, we share some of our thoughts and suggestions based on our experience with combining forecasts of the classical economic time series in this study.

6.1 Forecast Combinations in Applications

Regarding the ECB SPF application of forecast combinations, our findings are somewhat different from those of Genre *et al.* (2013). Genre *et al.* (2013) find that there is the strongest scope for improvement in the forecast accuracy in the case of the harmonised inflation over the equal weights benchmark as opposed to the cases of the real GDP growth and the unemployment rate regarding the variables, and that there is only a small scope for improvement in the case of 2 year horizon relatively to the case of 1 year horizon, regarding the individual forecast horizons. In contrast, the results of our testing in the table 5.4 suggest that the equal weights benchmark can be significantly outperformed equally well for all the ECB SPF macroeconomic variables and that the forecast combinations tend to outperform the equal weights in the case of the 2 year forecast horizon relatively more frequently than in the case of the 1 year forecast horizon. Furthermore, while Genre *et al.* (2013) emphasize their finding that the best combination methods vary across variables and horizons in the ECB SPF and so it is hard to make the case against the forecast combination puzzle, we, on the other hand, find that the Bates-Granger (4) forecast combination significantly improves upon the equal weights in all the datasets examined in ECB SPF application. Nevertheless, we find that the majority of forecast combinations examined in this study does not significantly outperform or are even significantly outperformed by the equal weights on the majority of datasets, which supports the

results of Genre *et al.* (2013) and Diebold & Shin (2017), that the equal weights (or simple average) indeed represents a strong benchmark for combinations of individual forecasts from the ECB SPF.

Our results from combining the individual realized volatility forecasts of the U.S. Treasury futures log-return application suggest that forecast combinations certainly have their use even in the financial applications. They prove useful in situations when one does not know a priori which of the individual forecast models will perform the best. Most of the methods can adjust the weights accordingly and outperform majority of the individual forecasts. In cases where there is some superior individual forecast, such as the HAR model in our application, some of the methods (e.g. the Market for Kernels, refer to tables 5.9 and 5.10) can readily assign most of the weight to the best performing individual forecast automatically. Our results are consistent with those of of Donaldson & Kamstra (1996) and Harrald & Kamstra (1997), who find that the ANN and EP-NN can outperform the simpler forecast combinations in forecasting the stock market daily volatility. However, regarding the futures in our application, we find that this result holds mostly for the larger training samples and longer forecast horizons of individual forecasts, while the neural network methods are significantly outperformed by nearly all of the other forecast combination methods in the opposite case. Nevertheless, the superiority of the performance of artificial neural networks in the case of large training samples and long horizons even to the best performing individual forecast (refer e.g. to the table 5.6) is an interesting phenomenon, likely attributable to the ability of the non-linear artificial neural networks to model and benefit from the interaction effects among the individual volatility forecasts discussed by Donaldson & Kamstra (1999).

6.2 Insights and Suggestions

Our work includes only empirical applications, with limited amount of datasets. We must therefore most certainly refrain from making any definite global conclusions about the performance of the studied forecast combinations. Nevertheless, our applications atleast partially cover combining of forecasts of the most classical of economic time series (real GDP growth, inflation, unemployment rate and volatility of financial asset log-returns). Moreover, to our knowledge, the range of the examined and compared forecast combination methods in a single study, is the largest up to date. Therefore, we feel that our results and suggestions might be useful to both academicians and practitioners applying the forecast combination techniques on a real data.

We believe that we have gathered some evidence against the forecast combination puzzle in the economic time series and that when the goal is not the inference but

rather obtaining as accurate forecasts as possible[1], it is worth to combine the forecasts beyond just the equal weights. In general, for combining macroeconomic forecasts as in the ECB SPF, where there are usually only small datasets available for training the combinations, we suggest using either some of the simple Bates-Granger optimal combining weights procedures, factor analytic (principal components) combinations or the artificial predictions markets. We cannot suggest using any of the other examined, usually more sophisticated, combining methods in such applications. For combining forecasts on large datasets in financial applications, we extend the pool of our suggested methods for the artificial neural networks and the BMA forecast combining methods as presented in this study.

We acknowledge that one could argue with our suggestions and object, that we have not examined the performance of the methods in all different possible parameter settings. For example, that we have not provided the methods with a sufficient number of iterations required for their full training or that we have not used enough cross-validating samples. And hence we cannot claim anything about the bad out-of-sample forecast performance of these methods. The one would be partially right. Nevertheless, bear in mind that we have mostly used the parameter settings either recommended or one of the best performing in a given empirical research. And that in practice, the amount of computational intensity required by the methods used often plays an important role. Consider e.g. the fast algorithmic trading on financial markets. So, we believe that reducing the maximum number of iterations allowed for each method on each rolling window down to a reasonable number is partially justified as the assessment of the practical relevance of the examined combination methods is one of our primary goals.

On the topic of which forecast combination method is generally the best, we do not believe there is any optimal combination method optimal in all practical situations, unless some very restrictive global theoretical assumptions are imposed. We deem that it is even impossible in principle, because there are some contradictory requirements (or tradeoffs) for its properties. Firstly, the method needs to be simple enough so it can work or its parameters can converge quickly on short datasets, while at the same time it needs to be flexible enough so that it can capture even more complex structures if necessary. See for example the case of the equal weights vs. the ANN in our applications. Secondly, the method needs to put heavier weights on the most recent performance of the individual forecasts, because there are more relevant to the current observation. However, at the same time, the method should work reliably and combine individual forecasts based on their overall performance in the sample, because the recent performance might be heavily influenced by the randomness of the underlying process. Take for example the case of the most success-

[1] Assuming the square loss function

ful method in our applications, the Bates-Granger (4), the Market for Kernels and the factor analytic (principal components) forecast combinations. What the Bates-Granger (4) and the Market for Kernels share in way is that they quickly transfer weights to the recently most accurate individual forecasts. This turns out to be a dominant strategy in both of our empirical applications. However, the factor analytic methods are based on entirely different principle, use equally the information from the whole training sample and still turn out to be among the best forecast combinations in both applications.

Finally, we would like to emphasize that the case has been made, atleast empirically, for the usefulness of the artificial prediction markets in the economic time series forecast combining, including the newly proposed Market for Kernels. The Market for Kernels shows to be a very simple, nonparametric, yet very effective method for combining forecasts, which performs significantly better or atleast comparably to the equal weights forecast in both of the empirical applications on almost all of the datasets. We wish our work would encourage further investigation of possible working artificial markets mechanisms applicable to economic time series forecasting.

Chapter 7

Conclusion

In this study, we have gathered and described a wide spectrum of forecast combination methods from the literature up to date and empirically assessed their (pseudo) out-of-sample forecasting performance in two of the most classical economic time series forecasting applications. The examined forecast combination methods were divided into classes: simple, factor analytic, shrinkage, bayesian model averaging, alternative and artificial prediction markets, roughly according to the principles of the common idea, complexity, relevance and time hierarchy. The first application in our work was combining the forecasts of individual contributors to the ECB quarterly Survey of Professional Forecasters. These included the individual forecasts of the real GDP growth, harmonised inflation and unemployment rate in 1 and 2 year horizons. Our second empirical application was combining the forecasts of the daily realized volatility of the U.S. Treasury futures log-returns. We examined the futures for the 2-Year, 5-Year and 10-Year Treasury Notes and the U.S. Treasury Bond with the 30-Year maturity. For combining, we used the 1, 5 and 22-steps ahead individual forecasts of the realized volatility from the commonly applied econometric models including the GARCH, HAR and several specifications of the VAR model. Each of these applications covers time series of a different length and nature and thus allowed us to inspect the behaviour and performance of the forecast combinations in various environments.

Our goal was to asses whether forecast combinations bring any improvements upon the original individual forecasts, compare the performance of different forecast combinations against each other and look for common patterns, from which any insights could be drawn that would be of use to both academicians and practitioners who wish to apply forecast combination techniques in order to improve accuracy of their forecasts. Next, we aimed to challenge the forecast combination puzzle, which states that it is hard in empirical applications to outperform the equal weights (simple average) forecast combination using other, more sophisticated methods. We

have also included in the set of evaluated methods the recently proposed artificial prediction markets methods c-APM (Constant) and c-APM (Q-learning), which is a class of machine learning methods inspired by the real prediction markets, and, to our knowledge, has not yet been applied to the problem of combining classical economic time series such as in our applications. Furthermore, we contribute to the pool of literature on the artificial prediction markets applicable to time series problems by introducing a new simple method called Market for Kernels and asses its forecast performance against the other forecast combination methods in both of the applications and against the individual forecasts in the financial application. For assessing the forecast performance we use the common measures of the forecast accuracy including RMSE, MAE and MAPE, and the test of equal forecast accuracy by Diebold & Mariano (2002).

Firstly, we found that the best individuals from ECB SPF were only rarely beat in accuracy measures by forecast combinations. Nevertheless, some of the simple, factor analytic and artificial prediction market forecasts repeatedly achieved better measures of forecast accuracy than the median individuals. In forecasting the realized volatility, most of the individual forecast combinations started to perform comparably to the best individual volatility forecasting model with the increasing sample length and forecast horizon or even surpassed it in case of e.g. ANN, EP-NN and the BMA forecast combinations. By comparing the performance of forecast combinations across all the datasets, we found that some of the simple, factor analytic and artificial prediction markets performed consistently well in both applications in relative to the other examined methods and we can suggest their use to practitioners in general economic times series forecasting problems. In applications, where there is a large amount of data for training the methods at disposal, we also suggest using the slightly more complex artificial neural networks and the BMA forecast combinations. Based on our data, we cannot recommend using the shrinkage methods and most of the methods in the alternative class as they have been shown to perform rather poorly in relative to the other methods. Further, we have found out, that a successful strategy in combining forecasts from both the ECB SPF and the forecasts of realized volatility of U.S. Treasury futures log-returns is to assign greatest weights to the most recently best performing individuals as this is the principal on which work the best method overall in our applications, the Bates-Granger (4) and the second best method, the newly proposed Market for Kernels. Regarding the forecast combination puzzle, we have found that the equal weights indeed are a strong benchmark as we could not reject the null hypothesis of equal forecast performance in favour of most of the forecast combinations on most of the ECB SPF datasets. Nevertheless, in contrast to some of the preceding literature, we found that some of the forecast combinations, namely the Bates-Granger (4), consistently significantly outperformed the

equal weights across all the variables and horizons. Moreover, most of the forecast combinations significantly outperformed the equals weights in our realized volatility application. Finally, the Market for Kernels method was found to either significantly outperform or atleast give a comparable performance to almost of all the other forecast combinations on the vast majority of datasets in both applications and also was found to significantly outperform all but the HAR model in forecasting the realized volatility.

We believe that this study has provided a useful guidance for anyone who wishes to apply forecast combination techniques in order to achieve as accurate forecasts as possible. We have also shown that the scope of useful methods for combining economic time series forecasts does not limit only to the traditional forecast combinations. The class of artificial prediction markets, including the proposed Market for Kernels, has been shown to combine forecasts comparably well to the best of the current benchmark forecast combinations. We hope that our findings will inspire future research of not only the artificial prediction markets, but also the non-traditional combinations of economic time series forecast in general.

Bibliography

Andersen, T. G., T. Bollerslev, F. X. Diebold, & H. Ebens (2001): "The distribution of realized stock return volatility." *Journal of Financial Economics* **61(1)**: pp. 43 – 76.

Andersen, T. G., T. Bollerslev, F. X. Diebold, & P. Labys (2003): "Modeling and forecasting realized volatility." *Econometrica* **71(2)**: pp. 579–625.

Armstrong, J. & F. Collopy (1992): "Error measures for generalizing about forecasting methods: Empirical comparisons." *International Journal of Forecasting* **8(1)**: pp. 69 – 80.

Barrow, D. K. & S. F. Crone (2016): "A comparison of adaboost algorithms for time series forecast combination." *International Journal of Forecasting* **32(4)**: pp. 1103 – 1119.

Bates, J. M. & C. W. Granger (1969): "The combination of forecasts." *Journal of the Operational Research Society* **20(4)**: pp. 451–468.

Bollerslev, T. (1986): "Generalized autoregressive conditional heteroskedasticity." *Journal of Econometrics* **31(3)**: pp. 307 – 327.

Bowles, C., R. Friz, V. Genre, G. Kenny, A. Meyler, & T. Rautanen (2007): "The ecb survey of professional forecasters (spf)-a review after eight years' experience." *ECB Occasional Paper No. 59. Available at SSRN: https://ssrn.com/abstract=967604* .

Buchen, T. & K. Wohlrabe (2011): "Forecasting with many predictors: Is boosting a viable alternative?" *Economics Letters* **113(1)**: pp. 16–18.

Bühlmann, P. & B. Yu (2003): "Boosting with the l2 loss." *Journal of the American Statistical Association* **98(462)**: pp. 324–339.

Burnham, K. P. & D. R. Anderson (2004): "Multimodel inference: Understanding aic and bic in model selection." *Sociological Methods & Research* **33(2)**: pp. 261–304.

Capistrán, C. & A. Timmermann (2009): "Forecast combination with entry and exit of experts." *Journal of Business & Economic Statistics* **27(4)**: pp. 428–440.

Chan, Y. L., J. H. Stock, & M. W. Watson (1999): "A dynamic factor model framework for forecast combination." *Spanish Economic Review* **1(2)**: pp. 91–121.

Claeskens, G., J. R. Magnus, A. L. Vasnev, & W. Wang (2016): "The forecast combination puzzle: A simple theoretical explanation." *International Journal of Forecasting* **32(3)**: pp. 754 – 762.

Conflitti, C., C. D. Mol, & D. Giannone (2015): "Optimal combination of survey forecasts." *International Journal of Forecasting* **31(4)**: pp. 1096 – 1103.

Cont, R. (2001): "Empirical properties of asset returns: stylized facts and statistical issues." *Quantitative Finance* **1(2)**: pp. 223–236.

Corsi, F. (2009): "A simple approximate long-memory model of realized volatility." *Journal of Financial Econometrics* **7(2)**: pp. 174–196.

Cowgill, B., J. Wolfers, & E. Zitzewitz (2009): "Using prediction markets to track information flows: Evidence from google." In "AMMA," p. 3.

De Menezes, L. M., D. W. Bunn, & J. W. Taylor (2000): "Review of guidelines for the use of combined forecasts." *European Journal of Operational Research* **120(1)**: pp. 190–204.

Demuth, H. B., M. H. Beale, O. De Jess, & M. T. Hagan (2014): *Neural Network Design*. USA: Martin Hagan, 2nd edition.

Diebold, F. X. (2015): "Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests." *Journal of Business & Economic Statistics* **33(1)**: pp. 1–1.

Diebold, F. X. & R. S. Mariano (2002): "Comparing predictive accuracy." *Journal of Business & Economic Statistics* **20(1)**: pp. 134–144.

Diebold, F. X. & P. Pauly (1990): "The use of prior information in forecast combination." *International Journal of Forecasting* **6(4)**: pp. 503–508.

Diebold, F. X. & M. Shin (2017): "Beating the simple average: Egalitarian lasso for combining economic forecasts." *PIER Working Paper 17-017*, Available at SSRN: https://ssrn.com/abstract=3032492.

Donaldson, R. G. & M. Kamstra (1996): "Forecast combining with neural networks." *Journal of Forecasting* **15(1)**: pp. 49–61.

Donaldson, R. G. & M. Kamstra (1999): "Neural network forecast combining with interaction effects." *Journal of the Franklin Institute* **336(2)**: pp. 227 – 236.

ECB (2018): "ECB Survey of Professional Forecasters." Retrieved at: http://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html.

Eklund, J. & S. Karlsson (2007): "Forecast combination and model averaging using predictive measures." *Econometric Reviews* **26(2-4)**: pp. 329–363.

Engle, R. F. (1982): "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation." *Econometrica* **50(4)**: pp. 987–1007.

Engle, R. F., C. W. Granger, & D. Kraft (1984): "Combining competing forecasts of inflation using a bivariate arch model." *Journal of economic dynamics and control* **8(2)**: pp. 151–165.

Fama, E. F. (1970): "Efficient capital markets: A review of theory and empirical work." *The Journal of Finance* **25(2)**: pp. 383–417.

Friedman, J. H. (2001): "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics* **29(5)**: pp. 1189–1232.

Garcia, J. A. (2003): "An introduction to the ecb's survey of professional forecasters." *ECB Occasional Paper No. 8. Available at SSRN: https://ssrn.com/abstract=748971* .

Genre, V., G. Kenny, A. Meyler, & A. Timmermann (2013): "Combining expert forecasts: Can anything beat the simple average?" *International Journal of Forecasting* **29(1)**: pp. 108–121.

George, E. I. & R. E. McCulloch (1997): "Approaches for bayesian variable selection." *Statistica Sinica* **7(2)**: pp. 339–373.

Gonçalves, S. & H. White (2004): "Maximum likelihood and the bootstrap for nonlinear dynamic models." *Journal of Econometrics* **119(1)**: pp. 199 – 219.

Granger, C. W. & Y. Jeon (2004): "Thick modeling." *Economic Modelling* **21(2)**: pp. 323–343.

Granger, C. W. & R. Ramanathan (1984): "Improved methods of combining forecasts." *Journal of Forecasting* **3(2)**: pp. 197–204.

Granger, C. W. J. & P. Newbold (1986): *Forecasting Economic Time Series*. Number 9780122951831 in Elsevier Monographs. Elsevier.

Green, P. J. (1995): "Reversible jump markov chain monte carlo computation and bayesian model determination." *Biometrika* **82(4)**: pp. 711–732.

Gu, S., B. T. Kelly, & D. Xiu (2018): "Empirical asset pricing via machine learning." *Chicago Booth Research Paper No. 18-04. Available at SSRN: https://ssrn.com/abstract=3159577 or http://dx.doi.org/10.2139/ssrn.3159577* .

Harrald, P. G. & M. Kamstra (1997): "Evolving artificial neural networks to combine financial forecasts." *IEEE Transactions on Evolutionary Computation* **1(1)**: pp. 40–52.

Hendry, D. F. & M. P. Clements (2004): "Pooling of forecasts." *The Econometrics Journal* **7(1)**: pp. 1–31.

Hu, J. & A. Storkey (2014): "Multi-period trading prediction markets with connections to machine learning." In "Proceedings of the 31st International Conference on Machine Learning," volume 32, pp. 1773–1781. Beijing, China.

Hyndman, R. J. & A. B. Koehler (2006): "Another look at measures of forecast accuracy." *International Journal of Forecasting* **22(4)**: pp. 679 – 688.

Inoue, A. & L. Kilian (2008): "How useful is bagging in forecasting economic time series? a case study of us consumer price inflation." *Journal of the American Statistical Association* **103(482)**: pp. 511–522.

Jacobson, T. & S. Karlsson (2004): "Finding good predictors for inflation: A bayesian model averaging approach." *Journal of Forecasting* **23(7)**: pp. 479–496.

Jahedpari, F., T. Rahwan, S. Hashemi, T. P. Michalak, M. De Vos, J. Padget, & W. L. Woon (2017): "Online prediction via continuous artificial prediction markets." *IEEE Intelligent Systems* **32(1)**: pp. 61–68.

Jeremy, S. & W. K. F. (2009): "A simple explanation of the forecast combination puzzle*." *Oxford Bulletin of Economics and Statistics* **71(3)**: pp. 331–355.

Lay, N. & A. Barbu (2010): "Supervised aggregation of classifiers using artificial prediction markets." In "Proceedings of the 27th International Conference on Machine Learning," pp. 591–598. Haifa, Israel.

Lay, N. & A. Barbu (2012): "The Artificial Regression Market." *ArXiv e-prints* .

MacQueen, J. *et al.* (1967): "Some methods for classification and analysis of multivariate observations." In "Proceedings of the fifth Berkeley symposium on mathematical statistics and probability," volume 1, pp. 281–297. Oakland, CA, USA.

McAleer, M. & M. C. Medeiros (2008): "Realized volatility: A review." *Econometric Reviews* **27(1-3)**: pp. 10–45.

Millin, J., K. Geras, & A. J. Storkey (2012): "Isoelastic agents and wealth updates in machine learning markets." In "Proceedings of the 29th International Conference on Machine Learning (ICML-12)," pp. 1815–1822. Edinburgh, Scotland, UK.

Newbold, P. & C. W. J. Granger (1974): "Experience with forecasting univariate time series and the combination of forecasts." *Journal of the Royal Statistical Society. Series A (General)* **137(2)**: pp. 131–165.

Newey, W. K. & K. D. West (1986): "A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix."

Pafka, S. & I. Kondor (2001): "Evaluating the riskmetrics methodology in measuring volatility and value-at-risk in financial markets." *Physica A: Statistical Mechanics and its Applications* **299(1)**: pp. 305 – 310. Application of Physics in Economic Modelling.

Politis, D., J. P. Romano, & M. Wolf (1997): "Subsampling for heteroskedastic time series." *Journal of Econometrics* **81(2)**: pp. 281 – 317.

Poon, S.-H. & C. W. Granger (2003): "Forecasting volatility in financial markets: A review." *Journal of Economic Literature* **41(2)**: pp. 478–539.

Raftery, A. E., D. Madigan, & J. A. Hoeting (1997): "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association* **92(437)**: pp. 179–191.

"Most Traded Futures" (2018): "CME Group Leading Products – Most Traded Futures and Options Contracts: Q1 2018." Retrieved at: https://www.cmegroup.com/education/files/cme-group-leading-products-2018-q1.pdf.

Said, S. E. & D. A. Dickey (1984): "Testing for unit roots in autoregressive-moving average models of unknown order." *Biometrika* **71(3)**: pp. 599–607.

Smith, J. & K. F. Wallis (2009): "A simple explanation of the forecast combination puzzle*." *Oxford Bulletin of Economics and Statistics* **71(3)**: pp. 331–355.

Stock, J. H. & M. W. Watson (1998a): "A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series." *Technical report*, National Bureau of Economic Research.

Stock, J. H. & M. W. Watson (1998b): "Diffusion indexes." *Working Paper 6702*, National Bureau of Economic Research.

Stock, J. H. & M. W. Watson (2004): "Combination forecasts of output growth in a seven-country data set." *Journal of Forecasting* **23(6)**: pp. 405–430.

Storkey, A. (2011): "Machine learning markets." In "Proceedings of the 14th International Conference on Artificial Intelligence and Statistics," volume 15, pp. 716–724. Fort Lauderdale, FL, USA.

Timmermann, A. (2006): "Forecast combinations." *Handbook of economic forecasting* **1**: pp. 135–196.

Wolfers, J. & E. Zitzewitz (2004): "Prediction markets." *Journal of Economic Perspectives* **18(2)**: pp. 107–126.

Yang, Y. (2004): "Combining forecasting procedures: some theoretical results." *Econometric Theory* **20(1)**: pp. 176–222.

Zhang, G., B. E. Patuwo, & M. Y. Hu (1998): "Forecasting with artificial neural networks:: The state of the art." *International Journal of Forecasting* **14(1)**: pp. 35 – 62.

Zou, H. & Y. Yang (2004): "Combining time series models for forecasting." *International journal of Forecasting* **20(1)**: pp. 69–84.

Appendix A

FV, TY, US - RVOL Figures and Tables

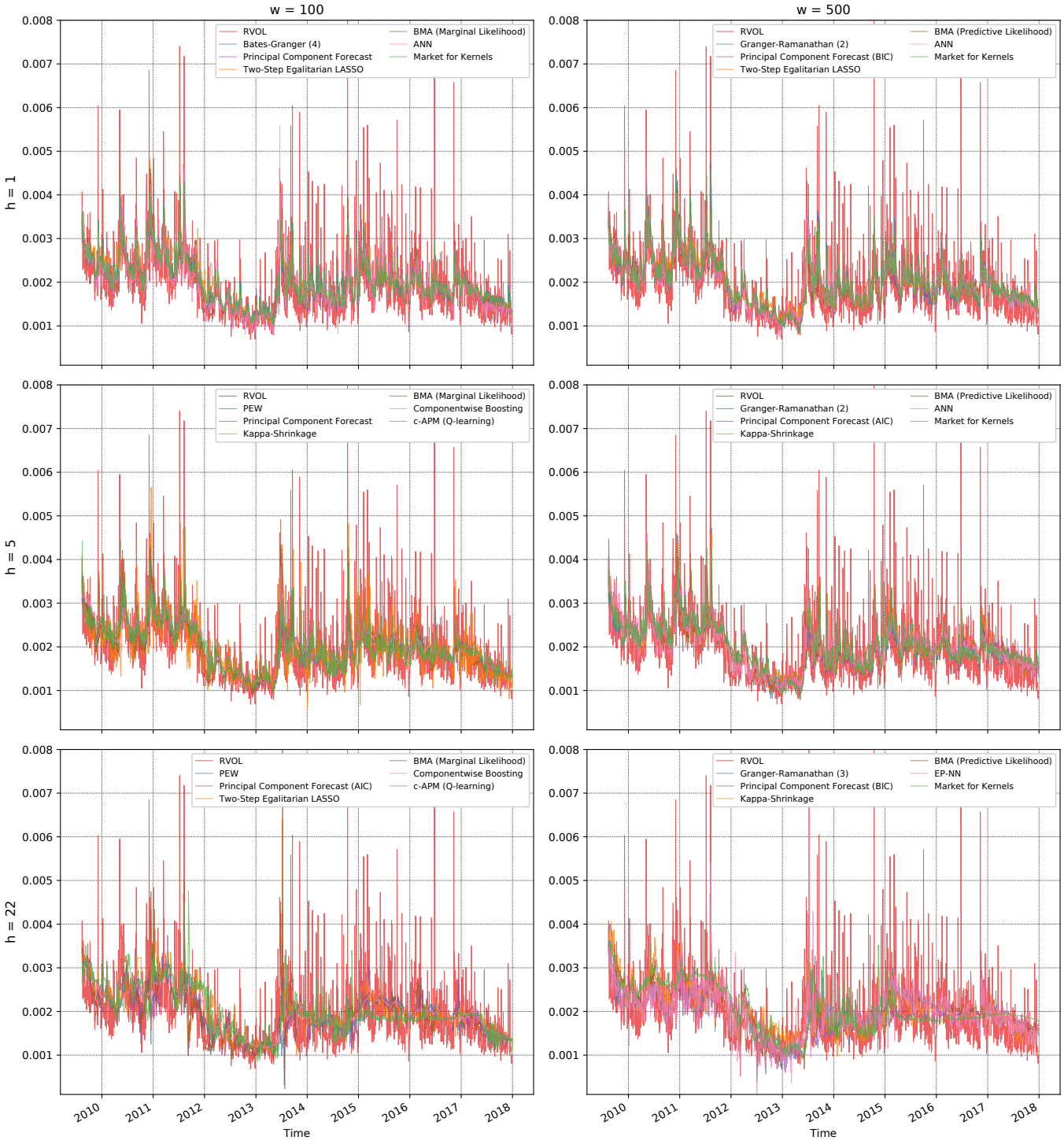

Figure A.1: Best combinations of h-steps-ahead forecasts of realized volatility of FV (5 Year) U.S. Treasury futures log-returns, trained on a rolling window of length w

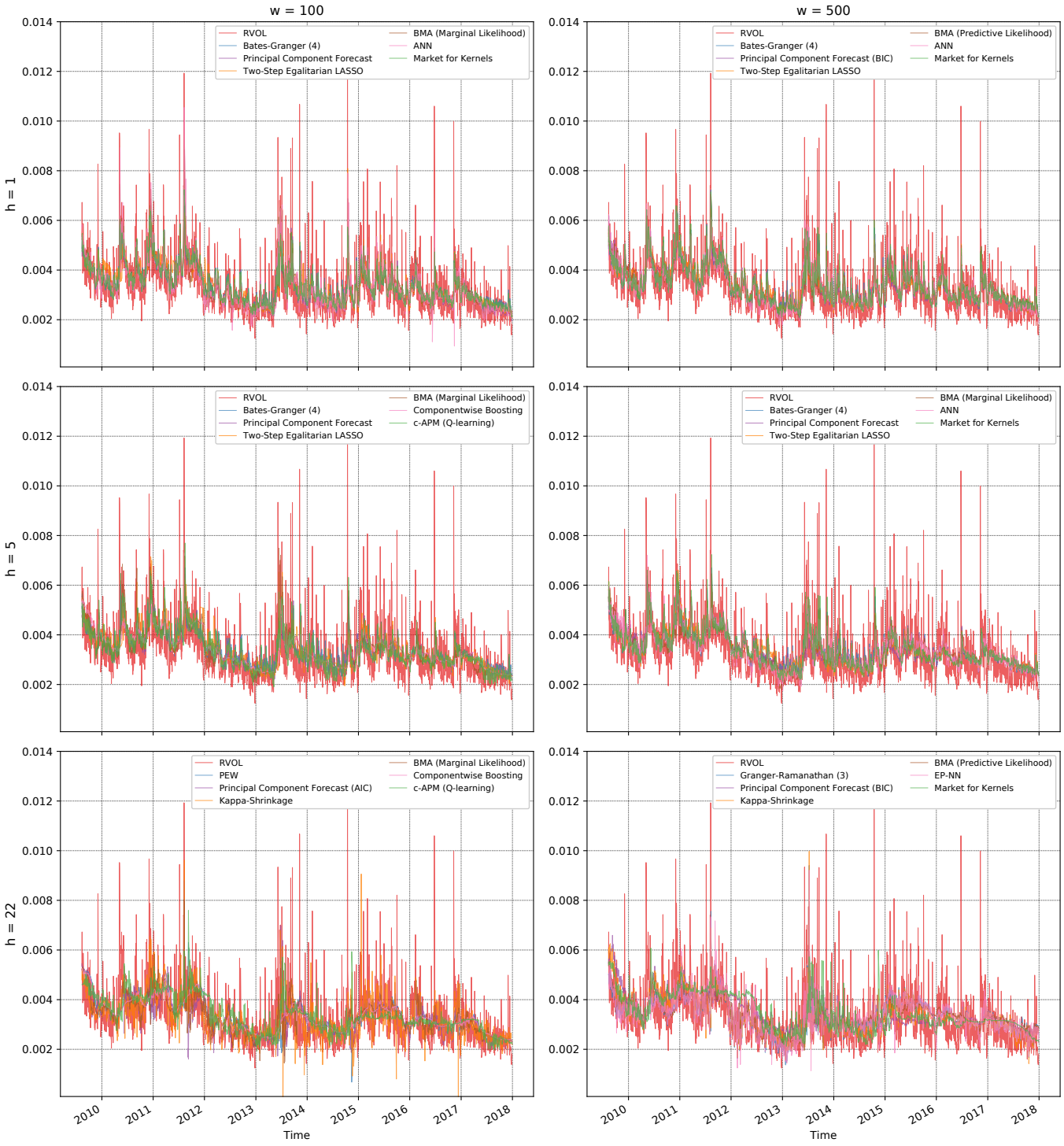

Figure A.2: Best combinations of h-steps-ahead forecasts of realized volatility of TY (10 Year) U.S. Treasury futures log-returns, trained on a rolling window of length w

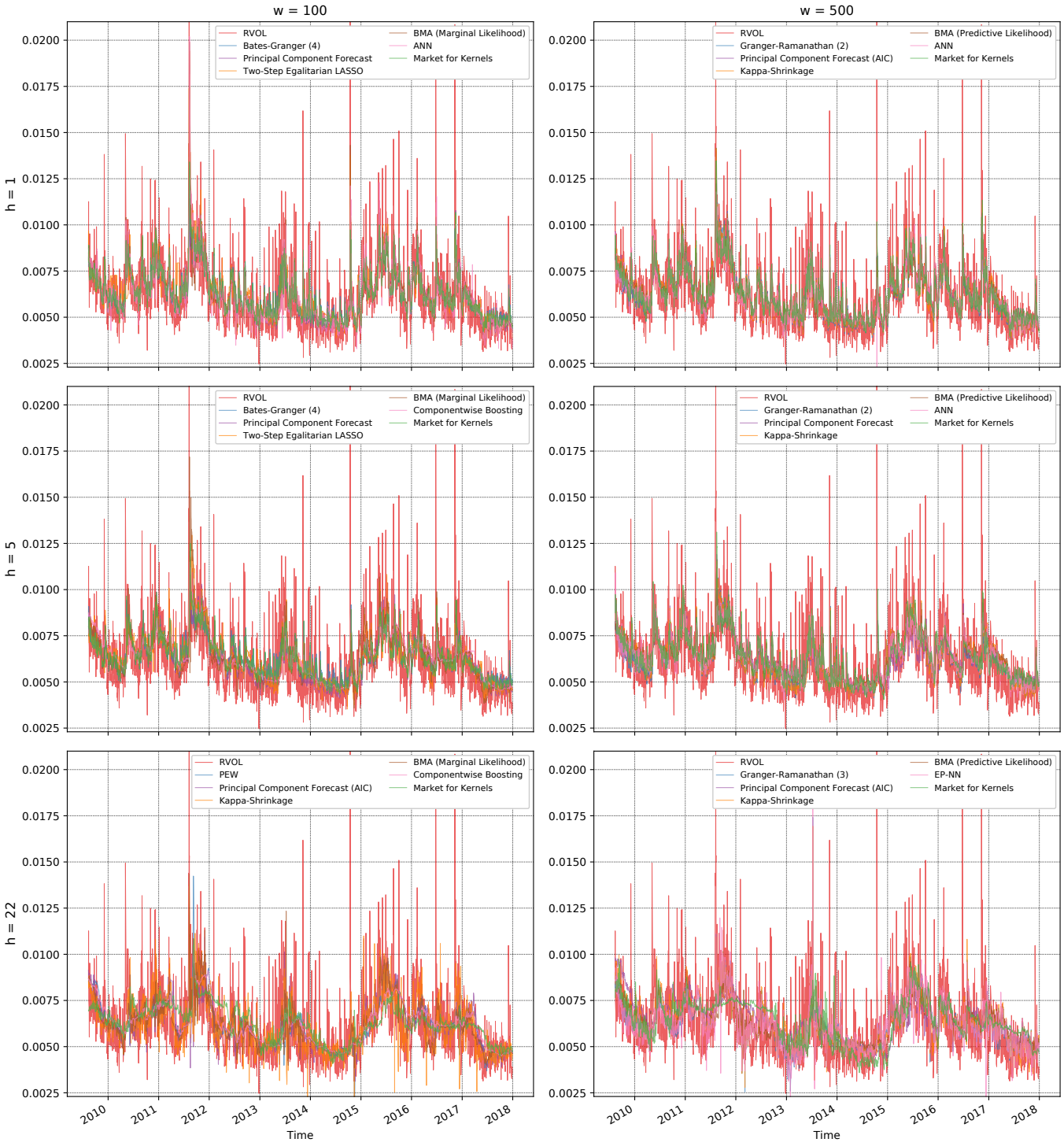

Figure A.3: Best combinations of h-steps-ahead forecasts of realized volatility of US (30 Year) U.S. Treasury futures log-returns, trained on a rolling window of length w

Table A.1: Performance of forecast combinations, trained on a rolling window of length w, of inidividual h-steps-ahead forecasts of realized volatility of log-returns of U.S. Treasury futures: FV (5 Year)

Class	Forecast Combination Method	h = 1						h = 5						h = 22					
		w = 100		w = 200		w = 500		w = 100		w = 200		w = 500		w = 100		w = 200		w = 500	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Simple	Equal Weights	7.61	22.35	7.43	22.49	6.47	23.26	8.43	26.15	8.13	26.32	7.03	27.25	9.75	33.31	9.45	33.66	7.98	35.23
	Bates-Granger (1)	7.55	21.32	7.38	21.51	6.42	22.23	8.33	24.75	8.07	25.14	6.97	26.26	9.36	29.74	9.18	30.98	7.87	34.00
	Bates-Granger (2)	7.54	21.62	7.38	21.71	6.42	21.92	8.38	25.80	8.11	25.97	6.97	26.18	9.77	33.45	9.49	34.14	7.94	35.10
	Bates-Granger (3)	7.58	21.92	7.41	22.08	6.45	22.84	8.39	25.57	8.10	25.84	7.00	26.85	9.57	31.83	9.33	32.55	7.93	34.73
	Bates-Granger (4)	7.52	20.94	7.34	21.03	6.39	21.60	8.21	23.53	7.92	23.64	6.80	24.28	9.06	27.49	8.82	27.78	7.29	28.81
	Bates-Granger (5)	7.56	21.96	7.38	22.09	6.45	22.82	8.40	25.80	8.09	25.95	6.98	26.82	9.78	33.42	9.47	33.73	7.97	35.30
	Granger-Ramanathan (1)	8.37	23.09	7.73	21.73	6.50	21.50	9.37	25.23	8.43	23.97	6.99	23.75	10.13	25.95	9.29	25.04	7.45	26.50
	Granger-Ramanathan (2)	8.32	22.31	7.67	21.07	6.45	20.55	9.16	24.52	8.40	23.43	6.94	22.53	10.16	25.72	9.19	24.60	7.41	24.95
	Granger-Ramanathan (3)	8.48	23.08	7.74	21.61	6.47	21.46	9.44	25.10	8.53	23.74	6.93	23.82	9.87	25.44	9.30	24.36	7.20	25.16
	AFTER	7.61	22.35	7.43	22.49	6.47	23.26	8.43	26.15	8.13	26.32	7.03	27.25	9.75	33.31	9.45	33.66	7.98	35.23
	Median Forecast	7.59	21.62	7.40	21.72	6.45	22.41	8.49	26.10	8.18	26.25	7.03	27.15	10.11	35.15	9.71	35.38	8.17	37.11
	Trimmed Mean Forecast	7.61	22.35	7.43	22.49	6.47	23.26	8.43	26.15	8.13	26.32	7.03	27.25	9.75	33.31	9.45	33.66	7.98	35.23
	PEW	7.58	20.99	7.41	20.87	6.44	20.65	8.17	23.03	7.97	23.41	7.05	23.99	8.59	24.24	8.61	25.69	8.09	29.38
Factor An.	Principal Component Forecast	7.53	20.01	7.38	20.43	6.46	21.42	8.19	21.97	8.00	22.65	6.97	24.10	8.85	24.40	8.82	25.73	7.78	29.15
	Principal Component Forecast (AIC)	7.61	20.48	7.46	20.55	6.43	20.60	8.34	23.20	8.01	23.07	6.90	23.19	8.47	23.32	8.42	24.62	7.59	27.25
	Principal Component Forecast (BIC)	7.60	20.10	7.40	20.17	6.42	20.60	8.35	22.78	8.04	22.98	6.93	23.60	8.52	23.77	8.42	24.65	7.57	26.99
Shrinkage	Empirical Bayes Estimator	16.53	41.47	15.01	33.04	10.63	36.03	19.43	45.08	14.77	43.48	8.98	29.04	46.04	74.51	70.61	63.33	11.01	32.84
	Kappa-Shrinkage	8.24	22.81	7.70	21.68	6.50	21.51	9.17	24.88	8.38	23.92	6.98	23.76	9.80	25.54	9.19	24.99	7.44	26.53
	Two-Step Egalitarian LASSO	8.00	22.74	7.52	22.16	6.47	21.91	9.20	25.93	8.14	25.38	7.01	26.29	9.29	26.93	8.99	28.67	7.89	31.67
BMA	BMA (Marginal Likelihood)	7.69	21.30	7.44	20.86	6.60	21.05	8.33	23.06	7.97	23.32	7.02	23.78	8.45	23.23	8.45	23.85	7.26	25.71
	BMA (Predictive Likelihood)	7.87	21.95	7.52	21.21	6.39	21.27	8.47	23.88	8.05	23.24	6.82	23.83	8.97	24.49	9.37	24.72	7.15	25.94
Alternative	ANN	7.77	21.40	7.54	21.24	6.42	21.27	8.59	23.75	8.05	23.01	6.84	23.85	8.80	23.82	8.52	24.52	7.43	27.43
	EP-NN	18.27	27.63	8.07	22.81	6.82	21.94	14.30	29.63	9.25	24.99	6.92	23.98	16.83	31.08	10.25	25.53	7.20	25.54
	Bagging	20.56	78.99	23.01	91.88	21.03	98.79	18.58	70.60	21.57	85.64	20.51	95.36	18.30	70.86	21.63	86.39	20.45	95.04
	Componentwise Boosting	7.80	22.87	7.59	23.57	6.65	26.24	8.26	24.10	8.05	25.41	7.05	28.32	8.49	24.37	8.60	27.10	7.72	31.40
	AdaBoost	8.92	28.22	9.08	30.79	8.86	38.42	8.88	28.05	8.98	30.93	8.98	38.98	8.74	27.13	8.88	30.06	8.87	38.42
APM	c-APM (Constant)	7.58	21.10	7.45	21.35	6.50	22.47	8.46	24.18	8.17	24.67	7.06	26.42	9.64	29.01	9.34	30.21	8.09	35.54
	c-APM (Q-learning)	7.65	20.72	7.52	21.17	6.49	22.23	8.38	22.82	8.18	23.24	7.01	25.57	9.42	26.48	9.07	27.62	8.10	32.90
	Market for Kernels	7.55	21.13	7.40	20.93	6.44	21.36	8.32	24.36	8.05	23.68	6.96	24.28	9.22	28.31	8.96	28.07	7.63	28.89
	Best Individual	7.51	20.61	7.34	20.63	6.40	20.99	8.32	22.85	8.08	22.85	6.90	22.86	9.44	27.24	9.34	27.33	7.72	26.71
	Median Individual	7.69	21.88	7.52	22.01	6.54	22.67	8.63	26.39	8.33	26.54	7.13	27.46	10.22	35.38	9.83	35.58	8.19	37.23
	Worst Individual	11.91	45.23	11.54	45.60	10.08	48.19	11.94	45.43	11.51	45.82	10.12	48.48	12.03	46.25	11.67	46.70	10.25	49.52

Note 1: The RMSE measure is scaled up by the order of 10^4.

Note 2: The MAE measure delivers similar message as the RMSE and so we omitted it from the table for space reasons.

Table A.2: Performance of forecast combinations, trained on a rolling window of length w, of inidividual h-steps-ahead forecasts of realized volatility of log-returns of U.S. Treasury futures: TY (10 Year)

Class	Forecast Combination Method	h = 1						h = 5						h = 22					
		w = 100		w = 200		w = 500		w = 100		w = 200		w = 500		w = 100		w = 200		w = 500	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Simple	Equal Weights	11.15	19.36	11.04	19.46	9.53	19.92	12.34	22.41	12.12	22.52	10.38	22.97	14.10	27.64	13.84	27.76	11.55	28.52
	Bates-Granger (1)	11.09	18.91	11.00	19.05	9.50	19.47	12.24	21.79	12.07	22.07	10.34	22.57	13.67	26.04	13.56	26.80	11.49	28.15
	Bates-Granger (2)	11.09	19.12	11.02	19.05	9.51	19.34	12.32	22.31	12.13	22.53	10.39	22.92	14.17	27.92	13.90	27.95	11.54	28.53
	Bates-Granger (3)	11.12	19.17	11.02	19.29	9.52	19.73	12.30	22.15	12.10	22.33	10.36	22.81	13.91	26.96	13.72	27.36	11.52	28.37
	Bates-Granger (4)	11.06	18.74	10.95	18.82	9.48	19.20	12.07	21.02	11.89	21.14	10.16	21.50	13.20	24.33	13.06	24.57	10.91	25.27
	Bates-Granger (5)	11.16	19.41	11.04	19.51	9.55	19.96	12.34	22.39	12.12	22.48	10.37	22.87	14.10	27.62	13.83	27.72	11.53	28.50
	Granger-Ramanathan (1)	12.55	21.13	11.35	19.54	9.67	19.14	13.57	23.09	12.43	21.80	10.54	21.35	13.84	22.15	13.57	22.36	10.96	22.52
	Granger-Ramanathan (2)	12.29	20.30	11.45	19.13	9.59	18.70	13.29	22.14	12.38	21.00	10.50	20.59	14.17	22.18	13.13	22.17	10.98	22.01
	Granger-Ramanathan (3)	13.00	21.08	11.42	19.42	9.64	19.31	13.62	23.01	12.50	21.79	10.58	22.07	13.23	22.08	13.76	21.83	10.82	22.58
	AFTER	11.15	19.36	11.04	19.46	9.53	19.92	12.34	22.41	12.12	22.52	10.38	22.97	14.10	27.64	13.84	27.76	11.55	28.52
	Median Forecast	11.20	19.15	11.07	19.24	9.56	19.66	12.54	22.71	12.28	22.77	10.44	23.19	14.68	29.29	14.27	29.21	11.74	29.95
	Trimmed Mean Forecast	11.15	19.36	11.04	19.46	9.53	19.92	12.34	22.41	12.12	22.52	10.38	22.97	14.10	27.64	13.84	27.76	11.55	28.52
	PEW	11.14	18.98	11.07	19.04	9.57	18.74	12.06	21.00	11.98	21.40	10.50	21.72	12.61	22.13	12.87	23.57	11.87	26.33
Factor An.	Principal Component Forecast	11.06	18.10	11.00	18.32	9.51	18.73	12.06	19.90	11.98	20.37	10.29	20.82	12.97	21.96	13.08	23.07	11.40	24.65
	Principal Component Forecast (AIC)	11.25	18.60	11.08	18.54	9.54	18.59	12.38	21.32	11.99	20.97	10.35	20.80	12.64	21.73	12.77	22.50	11.43	24.50
	Principal Component Forecast (BIC)	11.20	18.22	11.07	18.38	9.53	18.62	12.36	20.75	12.01	20.70	10.39	21.06	12.83	21.86	12.80	22.82	11.41	24.42
Shrinkage	Empirical Bayes Estimator	38.32	33.94	19.73	32.58	10.66	22.26	23.85	35.21	23.38	33.67	11.88	24.23	70.00	74.68	24.92	33.90	13.60	28.37
	Kappa-Shrinkage	12.36	20.81	11.32	19.49	9.66	19.14	13.30	22.71	12.38	21.72	10.53	21.34	13.46	21.81	13.43	22.28	10.94	22.52
	Two-Step Egalitarian LASSO	11.67	20.32	11.12	19.55	9.55	19.39	12.80	22.58	12.11	22.14	10.33	21.52	13.55	23.93	13.19	24.86	11.54	25.75
BMA	BMA (Marginal Likelihood)	11.31	19.33	11.06	18.81	9.54	18.76	12.19	21.13	11.94	21.26	10.32	21.21	12.33	21.03	12.42	21.96	10.82	23.33
	BMA (Predictive Likelihood)	11.48	19.65	11.16	19.19	9.48	19.09	12.56	22.09	12.19	21.35	10.23	21.61	12.91	21.81	13.51	22.32	10.63	22.90
Alternative	ANN	11.52	19.57	11.10	18.90	9.58	19.07	13.69	21.65	11.93	20.74	10.25	21.25	13.05	21.66	12.70	22.06	11.05	24.59
	EP-NN	25.85	25.27	14.24	20.92	9.82	19.65	52.92	29.62	16.13	23.11	10.55	22.05	22.92	26.95	19.27	22.99	10.87	23.09
	Bagging	31.76	77.88	37.13	92.79	35.15	99.06	31.13	75.25	36.86	91.24	34.76	98.16	28.75	68.02	32.67	76.79	30.81	81.43
	Componentwise Boosting	11.52	20.66	11.37	21.02	9.97	23.03	12.22	21.83	12.09	22.69	10.59	24.74	12.49	21.93	12.87	23.99	11.31	26.55
	AdaBoost	13.05	24.85	13.55	27.04	12.64	31.75	12.98	24.67	13.46	27.12	12.88	32.47	12.86	24.04	13.32	26.38	12.74	31.79
APM	c-APM (Constant)	11.14	18.91	11.12	19.07	9.61	19.71	12.54	22.29	12.29	22.50	10.51	23.38	14.17	27.18	14.15	27.98	11.69	29.80
	c-APM (Q-learning)	11.31	18.99	11.13	19.14	9.62	19.81	12.35	20.60	12.28	21.36	10.41	22.12	13.40	23.18	13.50	25.39	11.85	27.53
	Market for Kernels	11.11	18.72	11.04	18.65	9.50	19.00	12.20	21.38	12.00	21.04	10.34	21.27	13.34	24.68	13.44	25.72	11.49	25.19
	Best Individual	11.02	18.57	10.91	18.59	9.49	18.60	12.20	20.66	12.06	20.70	10.33	20.23	13.78	24.37	13.79	24.50	11.64	23.66
	Median Individual	11.35	19.47	11.25	19.49	9.68	19.84	12.68	22.99	12.43	23.02	10.54	23.42	14.83	29.54	14.39	29.42	11.83	30.11
	Worst Individual	16.66	35.01	16.39	35.10	13.57	35.82	16.72	35.16	16.39	35.26	13.62	36.01	16.92	35.77	16.63	35.88	13.77	36.71

Note 1: The RMSE measure is scaled up by the order of 10^4.

Note 2: The MAE measure delivers similar message as the RMSE and so we omitted it from the table for space reasons.

Table A.3: Performance of forecast combinations, trained on a rolling window of length w, of inidividual h-steps-ahead forecasts of realized volatility of log-returns of U.S. Treasury futures: US (30 Year)

Class	Forecast Combination Method	h = 1						h = 5						h = 22					
		w = 100		w = 200		w = 500		w = 100		w = 200		w = 500		w = 100		w = 200		w = 500	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Simple	Equal Weights	17.70	17.05	17.69	17.15	15.32	17.29	19.67	19.61	19.62	19.74	16.92	19.83	22.11	23.77	22.03	23.92	18.75	24.12
	Bates-Granger (1)	17.64	16.79	17.66	16.93	15.28	17.02	19.57	19.27	19.56	19.52	16.89	19.63	21.71	22.89	21.77	23.52	18.71	23.99
	Bates-Granger (2)	17.67	16.95	17.72	16.91	15.28	16.83	19.56	19.49	19.66	19.75	16.89	19.55	22.15	23.81	21.94	23.71	18.81	24.16
	Bates-Granger (3)	17.67	16.94	17.68	17.06	15.30	17.18	19.63	19.47	19.60	19.64	16.91	19.75	21.93	23.39	21.92	23.75	18.74	24.07
	Bates-Granger (4)	17.62	16.67	17.62	16.76	15.28	16.83	19.34	18.71	19.31	18.83	16.66	18.88	20.95	21.32	20.94	21.54	17.94	21.81
	Bates-Granger (5)	17.69	17.08	17.69	17.19	15.33	17.34	19.62	19.56	19.57	19.68	16.87	19.76	22.10	23.82	22.02	23.97	18.76	24.13
	Granger-Ramanathan (1)	19.84	18.91	18.42	17.64	15.37	16.59	22.04	20.98	20.17	20.02	16.99	18.60	21.93	20.28	20.96	19.93	18.10	20.95
	Granger-Ramanathan (2)	19.38	18.49	18.31	17.21	15.31	16.19	21.44	20.55	20.11	19.70	16.87	18.03	21.79	20.20	21.16	19.90	18.09	20.44
	Granger-Ramanathan (3)	20.23	19.13	18.58	17.49	15.35	16.53	22.32	20.95	20.32	19.89	17.01	18.92	21.99	20.22	21.01	20.01	17.93	20.48
	AFTER	17.70	17.05	17.69	17.15	15.32	17.29	19.67	19.61	19.62	19.74	16.92	19.83	22.11	23.77	22.03	23.92	18.75	24.12
	Median Forecast	17.88	16.94	17.88	17.03	15.42	17.12	20.08	19.93	20.00	20.02	17.09	20.06	23.00	25.15	22.80	25.13	19.18	25.24
	Trimmed Mean Forecast	17.70	17.05	17.69	17.15	15.32	17.29	19.67	19.61	19.62	19.74	16.92	19.83	22.11	23.77	22.03	23.92	18.75	24.12
	PEW	17.65	16.92	17.78	17.05	15.36	16.72	19.28	18.77	19.52	19.42	17.08	19.44	20.18	19.99	21.01	22.06	19.19	23.61
Factor An.	Principal Component Forecast	17.59	16.19	17.62	16.42	15.30	16.58	19.29	17.86	19.38	18.39	16.88	18.69	20.56	19.96	20.91	21.15	18.87	22.73
	Principal Component Forecast (AIC)	17.72	16.48	17.70	16.61	15.26	16.45	19.51	18.98	19.35	19.15	16.96	19.03	20.28	19.53	20.83	20.76	18.57	22.35
	Principal Component Forecast (BIC)	17.71	16.39	17.73	16.43	15.28	16.50	19.36	18.49	19.51	18.98	16.89	18.88	20.42	19.71	20.82	20.70	18.56	22.56
Shrinkage	Empirical Bayes Estimator	27.00	23.28	25.52	22.70	15.84	17.81	31.24	29.01	33.07	30.91	19.93	24.25	64.67	49.16	41.26	37.86	19.94	22.27
	Kappa-Shrinkage	19.52	18.62	18.35	17.57	15.36	16.58	21.49	20.57	20.06	19.91	16.97	18.59	21.41	19.95	20.80	19.86	18.07	20.94
	Two-Step Egalitarian LASSO	18.22	17.93	17.95	17.47	15.56	17.14	19.75	19.31	19.51	19.85	17.03	19.27	20.64	20.63	20.88	21.94	18.20	21.49
BMA	BMA (Marginal Likelihood)	17.99	17.11	17.89	16.89	15.29	16.70	19.34	18.95	19.43	19.20	16.92	19.21	19.42	18.95	20.44	19.98	18.02	21.34
	BMA (Predictive Likelihood)	18.25	17.81	17.83	16.96	15.26	16.63	20.87	19.95	19.71	19.29	16.83	19.26	21.14	19.74	21.81	20.65	17.51	20.59
Alternative	ANN	18.20	17.38	17.70	16.99	15.77	16.88	22.14	19.75	19.89	18.98	16.89	19.01	24.21	20.18	20.42	20.31	18.27	21.79
	EP-NN	55.97	23.36	26.54	18.90	16.76	17.33	27.29	23.91	22.35	21.03	17.32	19.54	34.96	24.28	29.34	21.47	17.99	20.39
	Bagging	54.97	76.86	64.93	92.84	64.43	98.48	52.67	70.82	65.26	92.45	64.72	99.06	51.41	69.67	58.69	79.28	57.43	86.51
	Componentwise Boosting	18.47	18.56	18.39	18.67	16.28	19.61	19.52	19.67	19.76	20.32	17.41	21.28	20.12	19.99	20.88	21.99	18.44	23.27
	AdaBoost	21.08	22.53	22.19	24.99	20.69	28.03	20.94	22.38	22.10	24.85	20.83	28.43	20.81	21.89	22.02	24.45	20.67	28.04
APM	c-APM (Constant)	17.97	16.84	18.02	17.02	15.47	17.13	20.02	19.77	19.89	19.91	17.07	20.06	22.56	24.02	22.47	24.56	19.16	25.24
	c-APM (Q-learning)	18.17	16.89	17.98	17.07	15.48	17.15	19.89	18.26	19.87	18.97	17.19	19.59	21.89	21.03	21.83	22.10	18.83	23.30
	Market for Kernels	17.73	16.58	17.78	16.63	15.30	16.69	19.54	18.72	19.48	18.53	16.96	18.91	21.49	21.54	21.39	21.92	18.79	22.34
	Best Individual	17.57	16.49	17.57	16.56	15.26	16.50	19.60	18.37	19.62	18.48	16.89	18.16	21.97	21.90	22.11	22.16	19.08	21.29
	Median Individual	18.13	17.26	18.14	17.35	15.61	17.25	20.28	20.22	20.20	20.28	17.26	20.05	23.16	25.41	22.93	25.35	19.31	25.37
	Worst Individual	25.45	29.71	25.30	29.74	21.14	29.80	25.55	29.82	25.38	29.85	21.22	29.92	25.91	30.27	25.76	30.31	21.41	30.37

Note 1: The RMSE measure is scaled up by the order of 10^4.

Note 2: The MAE measure delivers similar message as the RMSE and so we omitted it from the table for space reasons.