Universidade do Minho
Escola de Economia e Gestão

David Mendonça Pinho

Forecast comparison of volatility models and their combinations (FTSE100): a tied race

July 2020

Forecast comparison of volatility models and their combinations (FTSE100): a tied race
David Mendonça Pinho
UMinho | 2020

Universidade do Minho
Escola de Economia e Gestão

David Mendonça Pinho

Forecast comparison of volatility models and their combinations (FTSE100): a tied race

Master in Finance

Supervisor
Professor Doutor Nelson Areal

july 2020

Acknowledgements

First and foremost, I want to thank Professor Nelson Areal for his support. He was engaging to talk to, generous with his time, encouraging and kind – both as a professor and as an advisor. (I suspect he might very well be a saint.)

I am grateful to the University of Minho and my professors for the opportunities that were given to me and their focus on academic topics, which was very intellectually fulfilling to me. Lastly, I thank the ones that, regrettably, often end up getting the brunt of the unpleasantness despite being the most supportive: my close friends and family.

Statement of integrity

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Comparação de previsões de modelos de volatilidade e as suas combinações para o FTSE 100: uma corrida empatada

Resumo

Eu primeiro comparo 74 modelos que incluem os principais modelos simples, ARMA, GARCH, e HAR da literatura sobre o desempenho sem retrospetiva de modelos de volatilidade do dia seguinte. Para o índice FTSE100 no período de 2005-2010, todos os modelos HAR e GARCH e alguns modelos ARMA e de alisamento exponencial têm um desempenho similar. Eu, posteriormente, testo 176 combinações de modelos (o que significa que 250 modelos são comparados na totalidade) no período de 2007-2010, e observo que o desempenho médio tem menor variância e é sempre ligeiramente melhorado. Esta tendência não é observada com esquemas de ponderação complexos, que são baseados em regressões regularizadas (i.e., Lasso, Ridge e Elastic Net); e esta tendência é marginalmente maior quando se excluem modelos com mau desempenho passado e se ponderam igualmente as previsões dos modelos remanescentes (i.e., *trimming*). Mas, no geral, tal como é observado na literatura relevante, todos os modelos razoavelmente adequados tipicamente têm um desempenho idêntico, portanto a pesquisa passada parece ter exagerado ao reportar as melhorias geradas por novos modelos. Adicionalmente, existe um problema em que, de acordo com a literatura, mesmo em casos em que existem grandes melhorias no desempenho mensurado através das funções de perda usadas, isso raramente parece levar a melhorias em contextos económicos aplicados, tal como a gestão de risco e otimização de portefólios. Por causa disto, eu argumento que pesquisa subsequente tem de usar métricas diretamente relacionadas com estes contextos aplicados.

Palavras-chave: HAR, GARCH, combinações, regularização, volatilidade.

Forecast comparison of volatility models and their combinations for the FTSE 100: a tied race

Abstract

I first compare 74 models that include the main naïve, ARMA, GARCH, and HAR models from the volatility forecasting literature to assess their out-of-sample performance for day-ahead forecasts. For the FTSE100 index in the period of 2005-2010, all HAR and GARCH models and some ARMA and exponential smoothing models perform similarly to each other. I then test 176 model combinations (meaning that 250 models are compared in total) in the period of 2007-2010, and observe that the average performance has less variance and is always slightly improved. This tendency is not observed with complex weighting schemes, which are based on regularized regression (i.e., Lasso, Ridge and Elastic Net); and this tendency is marginally larger when excluding underperforming models and equal weighting the forecasts of the remaining models (i.e., trimming). But, overall, as observed in the relevant literature, all reasonably adequate models tend to have identical performance, so past research seems to have overstated the improvements generated by new models. An additional problem is that, according to the literature, even large performance gains with the loss functions used seem to rarely translate into improvements in economic applications, such as risk management and portfolio optimization. Because of this, I argue that subsequent research must use metrics directly related to these applications.

Keywords: HAR, GARCH, combinations, regularization, volatility.

Contents

List of Figures

List of Tables

1 Introduction

There is an extensive literature related to volatility forecasting that was propelled by the creation of the ARCH (Engle, 1982) and GARCH models (Bollerslev, 1986) and, later on, the use of intraday data to measure volatility with less noise (Andersen, Bollerslev, Diebold, and Ebens, 2001; Andersen, Bollerslev, Diebold, and Labys, 1999, 2003), which made HAR models (Corsi, 2009) the main focus of more recent research.

Both of these model families have spawned many variations in the literature, but it is not clear to what extent GARCH models are better than naïve models (Poon and Granger, 2005, 2003), or to what extent HAR models can be improved with other intraday volatility measures (Liu, Patton, and Sheppard, 2015), among other issues. At the same time, it is clear that larger, systematic tests or reviews (e.g., Hansen and Lunde, 2005; Liu et al., 2015; Poon and Granger, 2003) show that the performance of different models tends to be much closer than what a cursory read of the literature would suggest.

I empirically test a broad range of volatility models for one-day-ahead (out-of-sample) forecasts, similarly to the paper of Hansen and Lunde (2005) but with more model variety (74 individual models, and 176 model combinations). My main contributions to the literature include the comparison of naïve models, some of which perform comparably to GARCH and HAR models; the comparison of HAR models with more intraday measures, which was done to a more limited degree in Bollerslev, Patton, and Quaedvlieg (2016); and I combine these models with simple (e.g., equal weighting) and complex (i.e., regularized regression) weighting schemes, which was also done to a very limited degree for equity indexes (e.g., Becker and Clements, 2008). I also do a broad analysis of the literature related to significance testing because this an important topic to address when differences in performance are small, and multiple comparisons are being made.

I choose the individual models to test based mainly on two goals: I want to look at many model families; and, especially for HAR models, I select the ones that have the best performance in the literature. For model combinations, I mostly choose simple weighting schemes because that seems to work better in practice (Timmermann, 2006), although I also use regularized regression schemes, which are very similar to other approaches used in the literature. These choices allow me to observe the dispersion between the simplest models and the best models in the literature. With model combinations, I can also understand if inferior models still offer any improvements, despite their inferiority.

My main conclusion is that most models (some naïve models included) have identical performance. This is largely consistent with the findings of the literature, so I argue that I argue that researchers need to use metrics that are more closely aligned with improving risk management.

The remaining thesis is structured as follows. In section 2, I review the literature related to commonly used volatility models (section 2.1), the realised volatility measures they use (section 2.2), how they can be combined (section 2.3), and how their performance is evaluated – section 2.4 for the loss functions, and section 2.5 for the significance tests). I describe the data used, its sources, and some of its limitations on section 3. The methodology (section 4) presents the chosen loss functions (section 4.1), the individual models (section 4.2) and model combinations (section 4.3) tested, and the significance test (section 4.4) used to evaluate their performance. For the empirical results (section 5), I first present an overview of the main findings (section 5.1), most of them related to the differences between model families. In the following sections (from 5.2 to 5.3), for each model family, I analyse the subtler differences in performance, which tend to be less consistent or unsupported by the literature. The discussion (section 6) considers my results in the context of the broader literature: do other ways of improving forecasts perform better than model combinations? Do they lead to improvements in applied measures? And how should future research advance the knowledge in this area? Lastly, in section 7 I conclude by highlighting my main empirical findings and suggestions for future research.

Additionally, I provide suplementary material: the code I use for testing the models and analysing the results is available on a repository of my personal GitHub account (Pinho, 2020); Appendix A describes the general procedure I used for cleaning the data and some errors I found with it; Appendix B shows summary statistics for the volatility measures used (Table 7), and the correlation between the forecasts of model families (Table 8); and Appendix C contains the results presented in tabular form (Tables 9 through 17).

2 Literature review

In the literature about predicting volatility in financial markets, pure time-series models prevail. There have been attempts to use models that get closer to the underlying data-generating process (i.e., structural models), which use macroeconomic variables (Bollerslev, Tauchen, and Zhou, 2009) and measures of sentiment (Oliveira, Cortez, and Areal, 2017; Wang, Keswani, and Taylor, 2006), among others. But these attempts are ineffective, as they only manage to improve predictions by a small amount, at best. Contrarily, pure time-series models have been extremely effective – it is often possible to explain close to 80% of the variability in volatility (e.g., Andersen, Bollerslev, and Diebold, 2007; Corsi, Pirino, and Reno, 2010). The goal of such models is to accommodate the market's stylized facts, such as volatility clustering, excess kurtosis in the unconditional volatility process, or the leverage effect (Cont, 2001; Martens, Van Dijk, and De Pooter, 2009). These stylized facts have strong empirical and theoretical support, with some stylized facts like volatility clustering being observed in virtually every period, frequency, and financial market (Bollerslev, Chou, and Kroner, 1992; Bollerslev, Litvinova, and Tauchen, 2006). Importantly, we can assume these stylized facts will not be arbitraged away because volatility has no reason to stop being predictable – an efficient market does not remove the uncertainty that causes volatility. As such, the whole literature implicitly assumes that the current time-series structure related to volatility will persist.

2.1 Univariate time-series models

When forecasting volatility, "most of the seemingly chaotic nonlinearities work through the conditional variance," (Bollerslev et al., 1992, p. 23) which is why modelling it is the main focus of the research. Initially, the models conditioned volatility on squared innovations of the return process (Engle, 1982). The creation of the Generalized Autoregressive Heteroskedasticity (**GARCH**) model of (Bollerslev, 1986) soon revealed that the autoregressive component of volatility is more important. The GARCH conditions on past volatility in such a way that it simultaneously allows for serially correlated squared errors (i.e., volatility clustering), and serially uncorrelated returns (i.e., it allows you to model the return process as a random walk). Subsequent GARCH[1] models tried to incorporate other stylized facts, with two of them being emphasized: the leverage effect, which is the tendency for negative returns to increase volatility more than positive returns (Black, 1976; Nelson, 1991); and the tendency for large returns to affect volatility disproportionately more than

[1]When I write "GARCH" or "HAR", I am referring to the models from the family of GARCH or HAR models, respectively. When I use monospaced font (e.g., `GARCH(1,1)` or `HAR-RV`), I am referring to a specific model.

small returns (Bollerslev et al, 1992; Glosten, Jagannathan, and Runkle, 1993). In practice, most of the popular GARCH models react in similar ways to shocks in volatility (Hentschel, 1995) and have similar performance (Awartani and Corradi, 2005; Hansen and Lunde, 2005). These last two studies, in particular, show that for stock data and stock index data, the models that allow for leverage effects tend to be significantly better[2] than the `GARCH(1,1)`, though the improvements are still small: Hansen and Lunde (2005) report that, for most loss functions, the best models (out of 330) only reduce the loss by 5%-13% relative to the model with median performance; the model with median performance, in turn, consistently has identical performance to the `GARCH(1,1)`. Looking at a broader set of studies, Poon and Granger (2003) document that GARCH models are often surpassed by naïve models [3], but mostly in low-volatility periods.

Although GARCH models showed relatively unsatisfactory performance, they only started losing popularity after the development of realised variance (Andersen et al, 2001, 1999, 2003). Realised variance uses *intraday* returns instead of *daily* returns, which reduces the noise in the measurement of volatility (the noise comes, for example, from assuming that days with small daily returns are always days with low volatility). When there is less noise, volatility can be treated as an observable variable, which makes models much simpler and easier to estimate. One of those simpler models, the Heterogeneous Autoregressive (**HAR**) model of Corsi (2009), ended up becoming the standard in the literature. This model allows you to estimate its parameters with ordinary least squares (OLS); it seems to reduce the one-day-ahead mean squared error over the `GARCH(1,1)` models by roughly[4] 20%-60%; and it can replicate a large number of the already-mentioned stylized facts. It achieves this with parsimony because, unlike GARCH models, it does not need to model a long-memory process (where, for example, the volatility of one year ago can meaningfully affect today's volatility). Instead, it uses a short-memory process with multiple distributions, which makes the process empirically indistinguishable from a long-memory process (LeBaron et al, 2001).

Similarly to GARCH models, researchers developed new variants of HAR models that accommodate other features of financial markets. Some of the more common features have

[2] Using the test for superior predictive ability (Hansen, 2005), which controls for multiple comparisons. See more in section 2.5.

[3] Note that they consider "historical volatility" to include AR and multivariate VAR models, for example, so their results would not be as good if you applied my definition of "naïve model" expressed in section 4.

[4] There are very few studies directly comparing non-augmented HAR and GARCH models. Corsi (2004) compare them directly and finds a reduction in the MSE of almost 40%, with ARMA-type models having similar performance. Given that ARMA-type models are similar, we can extrapolate the results from other articles to obtain very rough estimates: Kambouroudis, McMillan, and Tsakou (2016) find that ARMA-type models, in relation to GARCH models, reduce the MSE by anywhere from 14% to 50%, which implies that differences between HAR and GARCH models can be larger; Andersen et al (2003) find a reduction ranging from 5% to 15%, but this is for exchange rate data, which shows that the chosen type of financial instrument has a substantial impact on the results.

to do with the leverage effect (Corsi and Reno, 2012; Patton and Sheppard, 2015), measurement error (Bollerslev et al., 2016), and jumps (Andersen et al., 2007; Corsi et al., 2010). Most variants provide small, inconsistent improvements over the base HAR model. There are two main exceptions to this. The first is the `HARQ` ("Quarticity HAR", which will be referred to as `Q-HAR` in this thesis) of Bollerslev et al. (2016) often gives sizable and consistent improvements in the forecast accuracy relative to the `HAR-RV` ("Realised Volatility HAR") model by incorporating information about the volatility of volatility. Using a large out-of-sample period, they find that the reduction in mean error is generally 3%-10% across forecast horizons, for both a stock index (S&P500) and individual stocks. The second exception is the `HAR-log`, which is more thoroughly explored in Buccheri and Corsi (2019); this model is identical to the `HAR-RV`, but it utilizes the logarithm of realised volatility instead of just realised volatility. The main improvement in performance over the `HARQ` happened during a period of higher volatility (2008-2013) for both the MSE and QLIKE loss functions.

2.2 Realised measures of volatility

Each HAR variant is associated with a type of realised measure that isolates a feature of the intraday time-series. For example, bipower variation (Barndorff-Nielsen and Shephard, 2004) tries to measure the intraday volatility without the influence of volatility jumps, which are periods of sudden and extreme increases in volatility caused by news releases or other events. These realised measures are then refined by other authors. One example for bipower variation is the *threshold* bipower variation (Corsi et al., 2010) that uses a jump-test to put more stringent requirements on what is considered a volatility jump.

Just like in the case of model variations, it is typical for the authors of new realised measures to claim that their contributions generated substantial improvements to model accuracy, but thorough empirical tests paint a different picture. One such test was conducted by Liu et al. (2015), where they use a very large number of measures with a data set that spans several asset classes. They conclude that the ubiquitous 5-minute realised variance measure is rarely (significantly) surpassed; and that the best measure across asset classes is the threshold realised variance, which is also very simple to calculate[5]. Patton and Sheppard (2009b) use a similar approach and find that the 5-minute realised volatility works well, but a simple combination of measures generally beats every individual measure; the accuracy improvements are small, however.

[5]To calculate the threshold realised variance, you exclude all the squared returns that are above a certain threshold. In the article, the threshold at day t is three times the average 5-minute intraday squared returns of day t.

2.3 Combining models

Earlier studies on forecasting volatility experimented with combinations of GARCH models and implied volatility ("IV" henceforth) (Donaldson and Kamstra, 2005; Lamoureux and Lastrapes, 1993; Martens and Zein, 2004; Pong, Shackleton, Taylor, and Xu, 2004) by including IV as an exogenous variable in the conditional volatility equation of GARCH models, and more recently Kambouroudis et al. (2016) do the same with realised volatility. Although this creates issues of multicollinearity – IV is very correlated with the conditional volatility (Becker, Clements, and White, 2007) – most results show that prediction accuracy can be substantially improved. These improvements tend to be concentrated in 1- to 10-day-ahead forecasts; for longer periods, the improvements get progressively smaller or negligible as the horizon increases in frequency to 22-day-ahead forecasts.

More recent studies find that trying to build a "super model" (Figlewski, 1997) is not the best approach. So instead of making one model that incorporates all sources of information and time-series properties, they combine several distinct individual models. With this approach, Becker and Clements (2008) forecast the *average* volatility for the following 22 days (rather than using step-ahead forecasts for individual days). When they separate implied volatility[6] from the remaining models, implied volatility is found to be much worse than model combinations, which are often the best models. There is a caveat, however: the VIX is among the best models when the accuracy is compared with a loss function that penalizes underpredictions more than overpredictions. So the VIX does worse largely because it has a risk-premium embedded in it, which causes a predictable upward bias (Chernov, 2007; Christensen and Prabhala, 1998).

Another small set of studies uses a different approach to combining models. In the case of Rapach and Strauss (2008) and Rapach, Strauss, and Wohar (2008), they first study to what extent there is evidence of structural breaks (i.e., abrupt changes in the structure of the time-series). They find strong evidence for it, so they test shorter lengths of estimation windows for GARCH models (among other things). Rapach and Strauss (2008) find that, for exchange rate data, using combinations of GARCH models with different estimation windows is better than choosing a single model/estimation window. This approach is not better because of large improvements to accuracy – the improvements are in the range of 2%-5%, typically. Instead, is it better because of the consistency: the improvements for combinations are much less dependent on the country which the data was tested on, and on forecast horizon and loss-function used. But the results tend to be mixed when we consider other studies. On one hand, Rapach et al. (2008) use an identical methodology and do not get nearly the same results, with the accuracy between individual models and combinations

[6] Measured with the model-free methodology discussed later, in the Methodology.

being almost indistinguishable. On the other hand, Yang, Tian, Chen, and Li (2017) use a HAR model instead of a GARCH one on Chinese agricultural futures, and the results are even better, with the reduction being larger than 10% over the base HAR model. The differences between studies are partly due to differences in the data – agricultural futures have more moderate structural breaks (Vivian and Wohar, 2012), and more prominent seasonality effects, which makes the base HAR a worse fit and creates an opportunity for models to exploit the additional structure.

There are adjacent works of literature to volatility forecasting that have explored model combinations more thoroughly. One minor instance of this happens in the equity risk premium literature. As an example, Rapach, Strauss, and Zhou (2010) first show that, in out-of-sample forecasts, all the methods they tested have worse accuracy than a simple historical average; but even when those forecasts are inferior by themselves, they become substantially better than the historical average when *combined*.

The vastest literature on forecast combinations is from Macroeconomics. Given that the discipline deals with related economic phenomena, we can also use the macroeconomic literature to understand which methods might work best. Clemen (1989) and Timmermann (2006) review the empirical literature and find the following stylized facts: good in-sample performance often leads to bad out-of-sample results; simple weighting schemes (e.g., equal weights for all models) outperform complex ones; and there are some small benefits when excluding the worst models ("trimming") or having time-varying weights[7]. All these facts are connected: they imply that simpler models and combinations work better. Even the more complex weighting schemes usually apply shrinkage towards equal weighting, with the improvements in prediction being very sensitive to the degree of shrinkage applied. There is one exception, however: trimming the worst forecasts *can* benefit prediction even when it is done aggressively. Aiolfi and Favero (2005), for instance, find that excluding around 80% of the models yields the best results. But these results should not generalise in a lot of settings because they are predicting stock returns, which are notoriously hard to

[7]Note that Timmermann (2006) has some errors in his treatment of the literature related to trimming, which is why I conclude slightly different things from his literature review. He often seems to confuse the meaning of "trimming" and "trimmed mean", though these have been used interchangeably (see Aiolfi and Favero (2005), for example). Trimming is a procedure where you exclude k% of the models with the worst past performance. Trimmed mean is a procedure that removes the most extreme forecasts in the cross-section of a day's forecasts; it does not look at past performance and is frequently symmetric (i.e., it excludes the k% forecasts that are largest and smallest, similarly to naïve outlier-removal methods). Here are, concretely, the mistakes Timmermann makes in the section about trimming (on page 183 and 184). He says that, according to Makridakis and Winkler (1983), "including very poor models in an equal weighted combination can substantially worsen forecasting performance," but that is not explicitly or implicitly stated by the authors of the article, and it cannot be taken from the results because the authors do not apply trimming. Stock and Watson (2004) also use trimmed mean and find that it is identical to using the mean or median forecast, but they do not use trimming. The only similar procedure to trimming that they perform is the "recent best" approach, picking the model with the best recent performance, but that is an extreme form of trimming, and not what Timmermann was referring to.

predict (Hou, Xue, and Zhang, 2017). This means that a larger share of the models might be completely non-informative, thereby making their exclusion worthwhile. Stock and Watson (2004) find that choosing the model with the best performance in the recent past often has good results, but it is extremely variable from country to country (similarly to other complex ways of making combinations), so the higher accuracy yielded by complex weighting schemes often comes at a cost.

Why is model simplicity better? It is partly because of sampling uncertainty: when two models have similar, yet distinct, performance, a relatively small sample will generate a noisy model ordering. By averaging the models' forecasts, we are ensuring that the best model is always weighted in, which will give us better accuracy if we have no way of knowing what the best model is; said another way, we tradeoff lower variance for higher bias. Strategies of combining models are themselves subject to the same tradeoff – combination strategies with simple weighting schemes have higher bias and lower variance.

In a lot of applied settings, there are much stronger reasons to prefer model simplicity. Though the sampling-based uncertainty still exists, social sciences have other worries – unobservable variables, shifting data-generating processes, and the adaptive nature of interacting agents are among the main ones. In this setting, most models are crude approximations, so it is unlikely that they can ever be fully correctly specified. Hence, the (apparent) best model in your sample is probably not taking full advantage of a given information set; the model structure might also not fit future data as well as it did in the past. Given our state of ignorance about the true data-generating process, we might get lower bias *and* variance by using model combinations.

Some studies show that more complex tools can improve prediction, however. The previously mentioned study of Becker and Clements (2008) and Yang et al. (2017) find that regression-based approaches outperform naïve weighting schemes, such as equally averaging all forecasts; within Macroeconomics, Rapach and Strauss (2010) also find success with more complex approaches; lastly, some competitors in the M4 competition (Makridakis, Spiliotis, and Assimakopoulos, 2020) use more complicated machine learning methods that generate improvements over simpler combinations, but that improvement is only around 5%, and is obtained in data sets with business applications that have a large seasonal component, which is not the case with equity markets' volatility. So, overall, there is more evidence in favour of simple combination schemes for volatility forecasting.

2.4 Loss functions

Every result reported so far is typically dependent on the volatility proxy and loss function used (see Brailsford and Faff (1996) or Brooks and Persand (2003) for extreme examples, or Poon and Granger (2003) more generally). For volatility proxies, it quickly be-

came clear that daily squared returns could be misleading in forecast comparisons (Andersen and Bollerslev, 1998). For loss functions, however, it was assumed that all would generate valid rankings. Thus, it was emphasized that you should use many different loss functions that are either more or less sensitive to outliers, or more or less sensitive to under/overpredictions (Brailsford and Faff, 1996; Hansen and Lunde, 2005).

But more recent developments showed that the majority of loss functions give invalid forecast rankings, particularly when using a noisy volatility proxy. The rankings are invalid in the sense that picking the model with lowest loss function will result in a biased variance estimate for finite samples (Hansen and Lunde, 2006; Patton, 2011). Patton (2011), in particular, derives a family of loss functions (dubbed "robust") that give accurate forecasts of the conditional variance, independently of the noise in volatility proxy, and of the data-generating process. Different robust[8] loss functions weight under/overpredictions differently, with the (commonly-used) mean squared error (MSE) placing no penalty for under or overpredictions; and the (also commonly-used) QLIKE placing a moderate penalty on underpredictions relative to overpredictions. As the penalty on over or underpredictions becomes higher, the statistical power of tests will tend to become smaller because you need a higher number of extreme observations to correctly estimate the true (asymptotic) value of the loss. But volatility models tend to underpredict large jumps in volatility, so the QLIKE is found to have the most power in statistical tests (Patton, 2011; Patton and Sheppard, 2009b).

Patton (2011) and Patton and Sheppard (2009a) argue that we should opt for robust loss functions in the absence of a well-defined economic application. It might seem that such an argument is an overstatement because the bias is very small when using less noisy proxies for volatility. In the case of the mean *absolute* error (MAE) of the variance, σ_t^2, there is a very small bias when using 5-minute realised variance – it minimizes $0.99\sigma_t^2$ (Patton, 2011). But this relies on optimistic assumptions that are often unmet in reality, such as assuming that the volatility process has no jumps. When Patton and Sheppard (2009a) use other assumptions for the data-generating process, the MAE minimizes $0.35\sigma_t^2$ in one extreme case. Thus, it is prudent to at least emphasize the results given by robust loss functions.

[8]Note that, in this case, the robustness refers to the asymptotic properties when varying the noise of the volatility proxy, or varying the data-generating process. These "robust" loss functions are, however, less robust in another sense: they are more impacted by outliers (e.g., when compared to the mean squared error, the mean absolute error is more robust to outliers, yet less robust to the noise in the volatility proxy or the data-generating process). In my case, the forecasts of most models are very correlated (Table 8), which means that outliers happen in the same days and have similar magnitudes. Hence, outliers often do not meaningfully impact the ranking of the models. Given this premise, I find it acceptable to tradeoff robustness against assumptions about the data-generating process for robustness against outliers.

2.5 Significance tests

Using the right loss functions gives us valid performance rankings asymptotically, but we still want to make inferences based on a sample. In finance, this is especially important because testing models on new data is often restricted by the passage of time (Lo and MacKinlay, 1990). For forecast comparisons, in particular, researchers achieve this goal by testing many models on the same data set, which creates the problem of data snooping[9] (i.e., it increases the probability that some model will significantly outperform all others by chance). Such a problem is not particularly acute when there is no selection bias (or other kinds of bias), because traditional statistical tests still give you valid results. For example, if you tested 100 models with the same performance and variance, and compared each pairwise difference in mean forecast error (with the test of Diebold and Mariano (2002), say), you would expect 5% of the differences to be statistically significant at the 5% level – just like if you made only one comparison. But researchers are seldom interested in knowing the result of an individual hypothesis in isolation. More commonly, they use (implicitly or not) tests of joint hypothesis to assess the strength of the evidence, which reduces the false-positive rate.[10] The corrections made for data snooping are not joint hypothesis tests, but they answer similar questions. The difference is that corrections for data snooping make the interpretation of individual hypothesis easier because, independently of the number of hypotheses created, their interpretation is always identical.

The Bonferroni correction (Dunn, 1961) and the Holm-Bonferroni test (Holm, 1979) are among the most notable procedures to guard against data snooping. They set a higher threshold for rejection of the null hypothesis, where that threshold increases as the number of hypotheses tested also increase. Both tests assume that p-values are independent, so they tend to be too conservative and reduce the statistical power of tests. For most empirical studies, this assumption of independence is not correct: the accuracy of volatility forecasting models will probably not be normally distributed and, more importantly, they will surely not be independent. This is because empirical studies test multiple models of the same family (e.g., HAR or GARCH models) that have very similar performance and similar p-values. Inspired by those tests, White (2000) and Romano and Wolf (2005) construct tests that estimate the dependence structure of the p-values, so they give higher statistical power[11], thereby reducing false-negatives.

All the tests previously mentioned use a researcher-specified benchmark. This can be a

[9] It is also called the problem of multiple comparisons, though that terminology is often associated with looking at multiple interactions in a model in the broader literature.

[10] This is true if you suppose that there are true and false hypothesis, and you are not sure which one is which. In that case, for instance, a simple rule like "only reject the null if the joint hypothesis test is also rejected" will reduce the number of false-positives.

[11] Note that the statistical power definition is different from test to test.

problem because it is possible to include inferior benchmarks and make the performance of the best model seem more significant. Hansen (2005) tries to correct for this with the test for superior predictive ability (SPA), but there are other problems associated with benchmarks. One of them is that a lot of settings have no natural benchmark; the other (larger) problem is that the benchmark may not be adequate to answer the researchers' question – a random walk model of realised volatility (where $\mathbb{E}[\text{RV}_t] = \text{RV}_{t-1}$) is often considered a natural benchmark (Andersen, Bollerslev, Christoffersen, and Diebold, 2005), but its performance will almost surely be inferior to any model that has an autoregressive component. Hence, we do not always get much information out of knowing if a model outperforms the benchmark.

The Model Confidence Set (MCS) of Hansen, Lunde, and Nason (2011) partially solves these issues because it does not require a benchmark. So, rather than assessing if models are significantly better than a benchmark, the MCS test compares all the models and yields a set of models (the so-called "set of superior models", or SSM) that are, for some level of significance, considered the "best". The inadequacies of the test arise out of the way that the test defines "best": the set of superior models is obtained by controlling the familywise error rate (FWER), meaning that it controls the rate at which *at least one* superior model is excluded from the set of superior models.

Hansen et al (2011, p. 454) argue that this objective is an appropriately sceptical goal because an informative sample "will result in a MCS that contains only the best model", and a less informative sample can fail to "distinguish between the models and may result in an MCS that contains several (or possibly all) models." But the problem is that the informativeness of a sample is not fully captured by the loss of the models. For instance, as I have previously argued, it is very common to see the rank of models change depending on the level of volatility. If we could incorporate this information, it would possibly allows us to know under what circumstances certain models might perform better or worse, so the conclusions would be less sample-dependent.

Another disadvantage of the MCS test, which is common to most hypotheses tests, is the lack of flexibility and interpretability when trying to test different, yet related, hypothesis. Some authors, for instance, try to prove the superiority of models with certain characteristics by counting how often each characteristic appears in the set of superior models (e.g., Liu et al (2015)). This is a very rough approximation to what the researchers want to know, as it treats outcomes as a binary event – the model is in or out of the SSM. It is also difficult to understand, for instance, if a model that is included 50% of the time is significantly better than another that is in the set 70% of the time (Gelman and Stern, 2006), or how the power of the test affects the results – are the models included in the set because they are almost surely equally good, or because of low power?

These disadvantages are apparent in the results of this study because the SSM includes

almost all models – even the ones that are blatantly inadequate. That said, I still consider it to be the most adequate significance test in the literature for the purposes of empirical model comparisons, although it mainly serves as a conservative benchmark.

3 Data

3.1 Period tested

I use data from the beginning of 2000 until the end of 2010 of the FTSE 100 index (ticker: FTSE100) and the FTSE 100 VIX index (ticker: VFTSEIX; Siriopoulos and Fassas, 2008) from Thomson Reuters' Datastream to construct daily returns and the implied volatility index, respectively. For the realised measures of volatility, I use OlsenData's intraday close-price data of the FTSE 100 index.

After cleaning the data as described in Appendix A, the sample has 2734 daily observations in total. For individual models, the initial estimation window ranges from the year 2000 to 2004 (5 years, 1249 observations[12]), and the models are trained on the 5-minute realised variance measure (RV, see section 3.2). For model combinations, the initial estimation window[13] ranges from 2005 to 2006 (2 years, 499 observations)[14], and they are trained on the forecasts or losses of the individual models. This leaves the comparison between all models' out-of-sample forecasts to range between 2007 and 2010 (4 years, 1004 observations). If a model needs to be estimated, it is refitted every day (this is implied in the future sections). For model combinations using regularization techniques, this procedure is slightly different (read more in the section 4.3.3).

In the body of this dissertation, I present results from 2005-2010, but the results from 2005-2006 are explored with less detail (Figure 1).

3.2 Volatility measures

Let the price of the FTSE 100 index, P_t, follow a continuous process expressed as a stochastic differential equation,

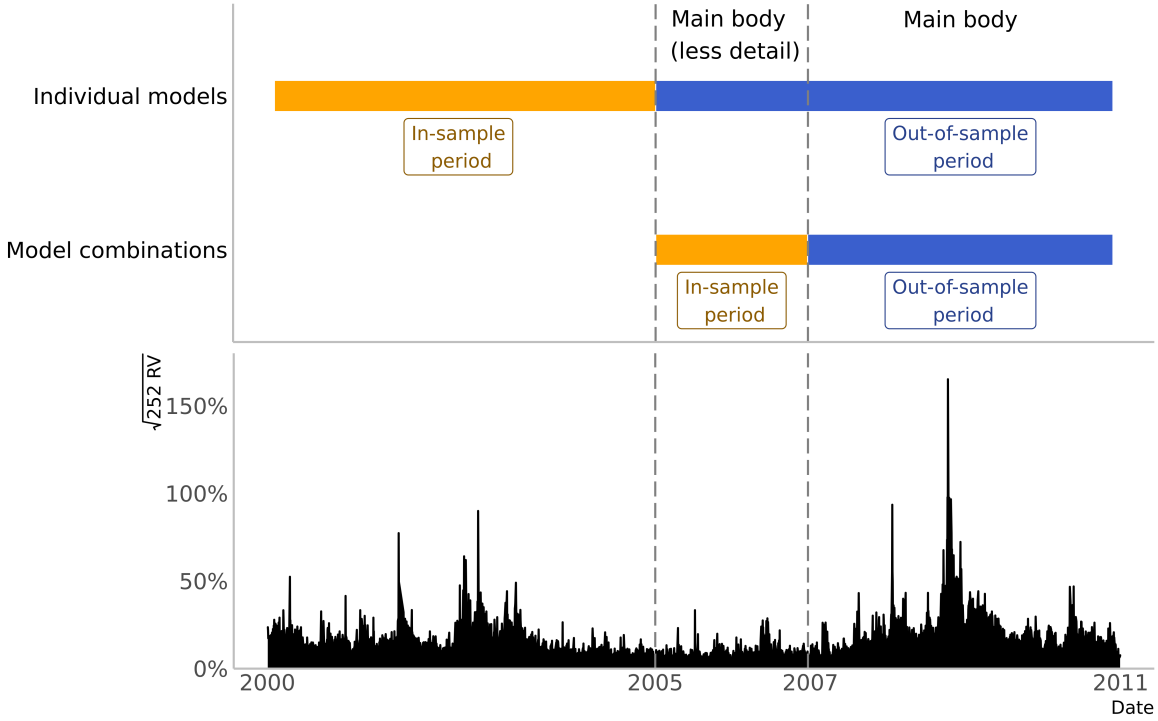
$$d\log(P_t) = \mu_t dt + \sigma_t dW_t + k_t dq_t. \tag{1}$$

[12] To train individual models, I use 5 years with an rolling expanding window instead of the typical rolling fixed estimation window of 1000 observations (e.g., Bollerslev et al. (2016); Hansen and Lunde (2005); Rapach and Strauss (2008, 2010)), and that seems to be sufficient for successfully estimating most models. At the same time, it seems that models with more parameters have some trouble with estimation (Bollerslev et al. (2016) mention such a case). My models are even harder to fit, so I use an expanding window of 1249 observations. In the studies that compare these choices, Rapach and Strauss (2008) find that GARCH models seem to perform better with fixed estimation windows of 1000 observations because the models do not have to explicitly account for structural breaks in the time series; for HAR models Buccheri and Corsi (2019) do not find any meaningful differences between fixed and expanding windows. So, in the end, there is a small tradeoff to using expanding windows.

[13] The procedure is somewhat different for model combinations with regularization techniques; read section 4.3.3.

[14] Other authors, such as Becker and Clements (2008), also use 500 observations to train combinations.

Figure 1: In-sample/out-of-sample split diagram (top) and realised variance (bottom)



The total sample starts in 01/01/2000 and ends in 31/12/2010. Individual models are trained on the realised variance (RV, see section 3.2). Model combinations are trained on the forecast errors of individual models. For model combinations using regularization techniques, this procedure is slightly different (read more in the section 4.3.3).

In the continuous part of the process ($\mu_t dt + \sigma_t dW_t$), μ_t is the drift process, σ_t is the instantaneous volatility process, and W_t is a standard Brownian motion[15]. In the part of the process that allows for jumps ($k_t dq_t$), k_t refers to the size of the jumps, and q_t is a counting process (typically a Poisson process) with a possibly time-varying intensity given by $\lambda_t dt$ (i.e., $\Pr[dq_t = 1] = \lambda_t dt$).

We are interested in finding the unobservable variance of this process, which is the quadratic variation, QV. The QV can be separated into the Integrated Variance (IV) and the residual jump component (JV):

$$QV_t \equiv \underbrace{\int_{t-1}^{t} \sigma_s^2 ds}_{IV_t} + \underbrace{\sum_{t-1<s\leqslant t} k_s^2}_{JV_t} . \tag{2}$$

[15]The process requires additional assumptions to allow for the existence of jumps; read (Andersen et al, 2007) for more theoretical background and references.

The volatility of volatility when the process is jump-free is given by the integrated quarticity, IQ_t, defined as

$$IQ_t \equiv \int_{t-1}^{t} \sigma_s^4 ds\,. \tag{3}$$

We only observe discrete prices ($P_{t,j}$) at some time interval, so the intraday returns ($r_{t,j}$) are calculated as

$$r_{t,j} \equiv \log \frac{P_{t,j}}{P_{t,j-1}},\ \text{for } j = 1, 2, ..., m \tag{4}$$

for day t in the j^{th} 5-minute interval[16], with the opening price being given when $j = 0$, and the closing when $j = m$.

Intraday returns would be calculated at the highest possible frequency if the intraday return process did not contain noise, as realised measures are often consistent estimators of their respective volatilities when the time between returns, Δ, tends towards 0:

$$QV_t = \operatorname*{plim}_{\Delta \to 0} RV_t\,; \tag{5}$$

$$IV_t = \operatorname*{plim}_{\Delta \to 0} TRV_t\,; \tag{6}$$

$$JV_t = \operatorname*{plim}_{\Delta \to 0} (RV_t - TRV_t)\,; \tag{7}$$

$$IQ_t = \operatorname*{plim}_{\Delta \to 0} RQ_t\,. \tag{8}$$

But we observe the return process with a lot of noise related to trading mechanisms (typically called microstructure noise, which is caused by the bid-ask spread bounce, periods of low liquidity, etc.). Because of this noise, I need to choose a return frequency that is high enough to measure the underlying volatility process, but low enough to not measure microstructure noise.

Provided that the underlying instrument is sufficiently liquid and tick-data is not available, most studies find that using 5-minute intraday returns is close to optimal (e.g., Andersen et al., 1999; Bandi and Russell, 2008; Liu et al., 2015; McAleer and Medeiros, 2008), so I use that periodicity. In the London Stock Exchange (LSE), the trading day has 8.5 hours (FTSE Russell, 2020) so the maximum amount of intraday returns, $1/\Delta$, is 102. In my data, the actual number of returns fluctuates from day to day, as the data can be very noisy and was cleaned to exclude certain observations.

[16] Note that the original unstructured data is in a higher frequency than 5-minutes, so $P_{t,j}$ is the price *at* time j or the last one available *before* time j.

I also use daily squared returns, r_t^2, as a proxy for volatility:

$$r_t^2 \equiv \left(\log \frac{P_{t,m}}{P_{t-1,m}} \right)^2, \tag{9}$$

where m is the last intraday price. Note that the information given by realised measures does not supersede the one given by squared returns – the realised measures do not include the close-to-open return, $\log \frac{P_{t,0}}{P_{t-1,m}}$.

Lastly, I use the FTSE 100 VIX index (ticker: VFTSEIX; Siriopoulos and Fassas, 2008) as the measure of (model-free) implied volatility. Let VFTSE_t be the value of the index; to create the implied daily variance, IV_t, we have:

$$\text{IV}_t = \left(\frac{\text{VFTSE}_t}{100\sqrt{252}} \right)^2. \tag{10}$$

Table 1 presents the formulas for the realised measures. In Appendix B, Table 7 shows summary statistics for all the volatility measures with the exception of implied volatility.

Table 1: Realised volatility measures

Realised measure	Authors	Variable	Definition
Realised variance	Andersen et al, 1999	RV_t	$\sum_{j=0}^{m} r_{t,j}^2$
Logarithmic Realised variance	Andersen et al, 2001	$RVlog_t$	$\log\left(\sum_{j=0}^{m} r_{t,j}^2\right)$
Truncated Realised variance	Liu et al, 2015; Mancini, 2009	TRV_t	$\sum_{j=0}^{m} r_{t,j}^2 \mathbb{I}_{\{r_{t,j}^2 > (3 \times \frac{RV_t}{m})\}}$
Jump	Barndorff-Nielsen and Shephard, 2004	J_t	$RV_t - TRV_t$
Realised quarticity	Barndorff-Nielsen and Shephard, 2002	RQ_t	$\frac{m}{3} \sum_{j=0}^{m} r_{t,j}^4$
Semi-variance (negative returns)	Barndorff-Nielsen, Kinnebrock, and Shephard, 2008	RV_t^-	$\sum_{j=0}^{m} r_{t,j}^2 \mathbb{I}_{\{r_{t,j} < 0\}}$
Semi-variance (positive returns)	Barndorff-Nielsen et al, 2008	RV_t^+	$\sum_{j=0}^{m} r_{t,j}^2 \mathbb{I}_{\{r_{t,j} > 0\}}$

As before, m represents the last observation of a given day, t, and $r_{t,j}$ is an intraday return as defined in equation (4).

4 Methodology

4.1 Loss functions

For the reasons mentioned in section 2.4, I only use the quasi-likelihood (QLIKE[17]) and mean squared error (MSE) loss functions that are defined for time series in the following way

$$\text{MSE}_t \equiv n^{-1} \sum_{i=0}^{n-1} (\text{RV}_{t-i} - \text{F}_{t-i})^2 ; \tag{11}$$

$$\text{QLIKE}_t \equiv n^{-1} \sum_{i=1}^{n-1} \left(\frac{\text{RV}_{t-i}}{\text{F}_{t-i}} - \log \frac{\text{RV}_{t-i}}{\text{F}_{t-i}} - 1 \right) ; \tag{12}$$

where RV and F are, respectively, the realised variance measure (Table 1) and the model's variance forecast; n is the number of observations. When $n = 1$, I reference the loss by squared error (SE_t) or QLIKE error (QLE_t).

4.2 Individual models

When comparing models, I want to reduce selection bias by testing as many models as possible, but that adds computational costs and, to some extent, makes the results harder to interpret. Given this constraint, I use a set of (loose) criteria to select the models from the literature:

- Model variety. Although a lot of models use the same information set, they use that information with a different structure. So even if a particular model is deemed inadequate in the literature, it might still provide benefits when combined with other models, especially when there are structural breaks (Pesaran and Timmermann, 2007).
- Model simplicity. As I argued in the Literature Review section, simpler models can often match or outperform relatively complex ones for individual models (for example, under calmer periods simple moving averages can do as well as GARCH models), and for model combinations (where equal weighting forecasts is hard to beat). Additionally, augmented individual models rarely provide any substantial benefit, so my study is more focused on testing whether models can be effectively combined (I explain this in more detail at the beginning of the section 4.3). Hence, most of the model complexity is present with model combinations.

[17] The QLIKE formula I used is the one defined in Liu et al. (2015) and Bollerslev et al. (2016), for example. It is slightly different from the commonly used formula and it makes the QLIKE always give positive values. That way, the ratio between the QLIKEs of models can be interpreted in a similar way to the ratio between the MSEs of models.

- Ease of estimation and implementation. If I do not find that a model has been created in the R language (R Core Team, 2020) as a function or package, or that it takes too much time to implement the model correctly and efficiently, I do not test that model. Additionally, I do not test some models that take a long time to fit, which happened with the FIGARCH model (Baillie, Bollerslev, and Mikkelsen, 1996).

- Past performance in the literature. I tend to not test models or alternative specifications that add complexity without improving forecast performance. Stochastic volatility models are an example of this because they are hard to fit and have bad performance compared to GARCH models (Poon and Granger, 2003).

These principles can create some selection bias, but, for this study, I think that is a very small issue for two reasons. Firstly, it is evident in the literature that the improvements over the simplest GARCH ($\texttt{GARCH}(1,1)$) and simplest HAR ($\texttt{HAR-RV}$) models are relatively small, so the differences between a biased and an unbiased sample also tend to be small. Secondly, a lot of papers empirically testing the performance of individual models were published before 2006 (with the exception of HAR models) . My out-of-sample period for model combinations starts after that year, so the bias would tend to be small.

Table 2: Number of individual models

Model family	Number of models
Exponential Smoothing	22
Rolling Average	16
GARCH	14
HAR	13
ARMA	6
Random Walk	2
Implied Volatility	1

When testing HAR models, I use specifications with a lot of parameters, and that goes (exceptionally) against the criterion of model simplicity. I do that because, to my knowledge, there is no broad empirical test with more complex specifications, unlike the case of GARCH models. I find it important to test them because it might be the case that one variable is positively correlated with volatility, while another is negatively correlated, so not adjusting for both variables makes both relationships seem weaker when they are tested independently – this is the "masked relationship" phenomenon (McElreath, 2020, p. 144-153).

In my study, for instance, it is plausible that volatility jumps lead to higher future volatility, while the noise in the volatility process can create the appearance of jumps.

Any negative forecast equal to or smaller than zero is set to half the previous day's forecast.

4.2.1 Naïve models

See Table 3 for the full list of models. As I mentioned in section 4.2, it is not uncommon to see simple models match the performance of GARCH models, so it is important to include them as a benchmark. To do this, I use some of the naïve models from Poon and Granger (2005, 2003)[18], although my definition of "naïve model" is more restrictive: it is any model that (conventionally) does not need to have its parameters estimated. In practice, this means that I do not consider models like the `AR(1)` as naïve.

The volatility proxy used, h, is either realised variance (RV) or squared returns (r^2). This allows me to have benchmarks for HAR and GARCH models, respectively.

Table 3: Naïve models

Full model name	Model name	Forecast
Random walk	$\texttt{RandWalk}_t^h$	h_{t-1}
Historical average	$\texttt{HistAvg}_t^h$	$n^{-1}\sum_{j=1}^{n} h_{t-j}$
Rolling average	$\texttt{RollAvg}(n)_t^h$	$n^{-1}\sum_{j=1}^{n} h_{t-j}$
Exponential smoothing	$\texttt{ExpSmooth}(\beta)_t^h$	$(1-\beta)h_{t-1} + \beta F_{t-1},\ 0 \leqslant \beta \leqslant 1$

h is the volatility proxy used – RV or r^2. For the historical average, n is equal to the cumulative number of observations at time t. For the rolling average, n corresponds to a fixed number of days that can take the value of $\{1, 5, 21, 63, 126, 252, 504, 1008\}$, meaning that I test 8 different models. For the exponential smoothing model, β can take the values of $\{0.05, 0.1, 0.2, ..., 0.9, 0.95\}$, so I test 11 different models. To initiate the exponential smoothing model, I set the initial value of the volatility proxy ($\texttt{ExpSmooth}(\beta)_1^h$) equal to the variance of that volatility proxy in the initial estimation period, which has 1231 observations (i.e., $\texttt{ExpSmooth}(\beta)_1^h = (n-1)^{-1}\sum_{t=0}^{n}(h_t - \bar{h})^2$). Similarly to GARCH models, a higher β parameter leads to more persistence in the volatility process.

[18] Note that some of the names and notation are different. This happens for naïve and all subsequent models presented.

4.2.2 ARMA models

The $\texttt{ARMA}^h(p, q)$ model is given by:

$$h_t = \mu + \sum_{i=0}^{p} \alpha_i \varepsilon_{t-i} + \sum_{i=0}^{q} \beta_i h_{t-i} + \varepsilon_t, \tag{13}$$

where h is the volatility proxy used (realised variance, RV, or squared returns, r^2), and ε is the residual term. I test all the models that are possible construct with the parameters p and q being 0 or 1, meaning that the `AR(1)` and `MA(1)` are both included in this category.

Much like HAR models, the process has short memory and uses the volatility proxy directly. This is not optimal for squared returns because ARMA models assume that the observed volatility (r_t^2) is equal to the conditional variance (σ_t^2), so very low values of r_t^2 are assumed to be low-volatility days. This issue does not exist nearly to the same extent with realised variance – in Table 7 from Appendix B, for the 2007-2010 period, the mean realised variance (2.2e−04) is close to the mean squared return (2.6e−04), but the realised variance at the 5^{th} percentile (2.7e−05) is more than fifty times higher than the squared returns at the 5^{th} percentile (4.3e−07). That said, I still test ARMA models because they combine features from HAR and GARCH models. Like HAR models, they treat volatility as an observable process; like GARCH models, they are not purely autoregressive processes, so large unexpected shocks impact the conditional volatility through the residuals, ε_{t-i}. Both of these things combined allow the ARMA models to possibly react quicker to sudden increases in the volatility, especially when compared to GARCH models.

Another alternative to ARMA models would be to use AR(F)IMA models, which can create long memory. They are, however, considerably harder to estimate, and they seem to have identical performance to HAR (Buccheri and Corsi, 2019; Corsi, 2009) and ARMA models (Becker and Clements, 2008; Kambouroudis et al., 2016; Pong et al., 2004). I had initially tested ARIMA models, and they do seem to have superior performance in some circumstances, but I do not have a good reason to include them.

Estimation and application

The ARMA models are tested with the "forecast" R package (Hyndman, Athanasopoulos, Bergmeir, Caceres, Chhay, O'Hara-Wild, Petropoulos, Razbash, and Wang, 2020). The parameters are estimated with maximum likelihood.

4.2.3 Univariate HAR models

Corsi (2009) treats volatility as an observable variable with the Heterogeneous Autoregressive model, `HAR-RV`:

$$\texttt{HAR-RV}_t = \alpha + \beta_1 RV_{t-1} + \beta_2 RV_{[t-1,t-5]} + \beta_3 RV_{[t-1,t-22]} + \varepsilon_t \,, \tag{14}$$

where RV is the realised variance (Table 1), α is the constant term, ε is the residual, and the subscript in $RV_{[t-1,t-T]}$ represents the average of (in this case) RV over a given period:

$$RV_{[t-1,t-T]} \equiv T^{-1} \sum_{i=1}^{T} RV_{t-i} \,. \tag{15}$$

There are three volatility components – the average daily ($T = 1$), weekly ($T = 5$), and monthly ($T = 22$) realised variances. Although it is possible to use more or fewer components, with or without time-variation, I do not test them because the performance improvements are negligible – Buccheri and Corsi (2019) use time-varying lags in their Score-HAR model, and the performance is seldom better than the HARlog; and Audrino and Knaus (2016) show that the specific lags used are not of great importance, so long as the lags are relatively small (i.e., smaller than 22 days).

In theory, the `HAR-RV` can be thought of as representing the heterogeneous opinions of market participants (Müller, Dacorogna, Davé, Pictet, Olsen, and Ward, 1993), where the volatility of each time horizon (daily, weekly, or monthly) is a proxy for the way those market participants react to new events – a trader looking to buy a stock without paying a large bid-ask spread might care about the daily volatility, while another trader looking to rebalance their portfolio might find the monthly volatility more relevant because it is more persistent.

In the alternative HAR specifications that I test, I distinguish the daily (D_{t-1}), weekly (W_{t-1}), and monthly (M_{t-1}) components of volatility, where each respective HAR model follows the process given by

$$HAR_t \equiv \alpha + D_{t-1} + W_{t-1} + M_{t-1} + \varepsilon_t \,, \tag{16}$$

where Table 4. Their forecasts (F_t) are similarly given by

$$F_t = \alpha + D_{t-1} + W_{t-1} + M_{t-1} \,, \tag{17}$$

except for the `HAR-RVlog` that is given by

$$F_t = \exp\left\{ \alpha + D_{t-1} + W_{t-1} + M_{t-1} + \frac{\sigma_\varepsilon^2}{2} \right\} \,, \tag{18}$$

where $\sigma^2_\varepsilon = \sum_{i=1}^{n} \frac{(\varepsilon_{t-i} - \bar{\varepsilon})^2}{n-1}$, which gives the variance of all past residuals (ε).

Not all volatility components change with the different realised measures used. Everything else constant, models with quarticity, `Q`, only change the daily component[19] by adjusting for the volatility of volatility (RQ, the realised quarticity) of the previous day. If RQ_{t-1} is high, there were likely periods of low liquidity, sudden trading activity, or other microstructure noise that in the short-term make prices strongly mean-reverting (Andersen et al, 1999). This behaviour is not relevant to the ultimate objective of measuring volatility: knowing how much prices might move. Thus, models with `Q` (typically) assume that the volatility is lower when RQ_{t-1} is higher.

In models with semi-variance, `S`, only the daily component changes by replacing RV_t with RV_t^+ and RV_t^- in an attempt to account for the leverage effect[20].

The models with the continuous measure, `C`, all volatility components use truncated realised variance, TRV (see Table 1), except in the models with semi-variance; in that case, only the weekly and monthly components use TRV. This is a measure of realised volatility that is jump-robust, similarly to the commonly used bipower variation, but it requires fewer assumptions about the volatility process and it is a more efficient estimator, in theory (Mancini, 2009) and practice (Liu et al, 2015). The models with the jump component, `J` (which are always accompanied by the continuous component[21]) always use jump realised measure ($J_t = RV_t - TRV_t$) in the daily, weekly, and monthly components. The jump measure is usually found to have almost no impact on volatility (e.g., Andersen et al (2007)), but this changes when using better jump-robust measures[22] (Corsi et al, 2010).

Lastly, I also test the `HAR-RVlog` because it was shown to perform better than a lot of other specifications, including the `Q-HAR` of Bollerslev et al (2016). To my knowledge, there are only two recent articles where this is shown more prominently. The first is by Buccheri and Corsi (2019). Even though the authors are arguing in favour of their more complex specifications, the `HAR-RVlog` has almost identical performance to those specifications in every sub-period. Similarly, Clements and Preve (2019) also partly argue in favour of estimating HAR models in alternative ways to ordinary least squares, but the `HAR-RVlog` has similar benefits of improving performance.

[19] This is sufficient for making one-step-ahead forecasts according to (Bollerslev et al, 2016).

[20] Corsi and Reno (2012) use another HAR model that accounts for the leverage effect, but they use daily squared returns instead of the semi-variance. That model gave me extremely variable estimates, a lot of them negative, so it looks to be a strictly inferior alternative to the model I use here.

[21] This is done because models with `RV-J` and models with `C-J` have very similar information, so it was not uncommon for the forecasts to be indistinguishable.

[22] This is not proven to be the case for TRV, but that seems to be a more accurate measure than what (Corsi et al, 2010) uses. So I expect that the TRV measure similar benefits to the ones described in (Corsi et al, 2010).

Table 4: All HAR models tested

Model name	D_{t-1}	W_{t-1}	M_{t-1}
`HAR-RV`	$\beta_1 RV_{t-1}$	$\beta_2 RV_{[t-1,t-5]}$	$\beta_3 RV_{[t-1,t-22]}$
`HAR-RVlog`	$\beta_1 RVlog_{t-1}$	$\beta_2 RVlog_{[t-1,t-5]}$	$\beta_3 RVlog_{[t-1,t-22]}$
`HAR-C`	$\beta_1 TRV_{t-1}$	$\beta_2 TRV_{[t-1,t-5]}$	$\beta_3 TRV_{[t-1,t-22]}$
`HAR-C-J`	$\beta_1 TRV_{t-1} + \beta_2 J_{t-1}$	$\beta_3 TRV_{[t-1,t-5]} + \beta_4 J_{[t-1,t-5]}$	$\beta_5 TRV_{[t-1,t-22]} + \beta_6 J_{[t-1,t-22]}$
`S-HAR-RV`	$\beta_1 RV^+_{t-1} + \beta_2 RV^-_{t-1}$	$\beta_3 RV_{[t-1,t-5]}$	$\beta_4 RV_{[t-1,t-22]}$
`S-HAR-C`	$\beta_1 RV^+_{t-1} + \beta_2 RV^-_{t-1}$	$\beta_3 TRV_{[t-1,t-5]}$	$\beta_4 TRV_{[t-1,t-22]}$
`S-HAR-C-J`	$\beta_1 RV^+_{t-1} + \beta_2 RV^-_{t-1} + \beta_3 J_{t-1}$	$\beta_4 TRV_{[t-1,t-5]} + \beta_5 J_{[t-1,t-5]}$	$\beta_6 TRV_{[t-1,t-22]} + \beta_7 J_{[t-1,t-22]}$
`Q-HAR-RV`	$(\beta_1 + \beta_2 RQ^{1/2}_{t-1}) RV_{t-1}$	$\beta_3 RV_{[t-1,t-5]}$	$\beta_4 RV_{[t-1,t-22]}$
`Q-HAR-C`	$(\beta_1 + \beta_2 RQ^{1/2}_{t-1}) TRV_{t-1}$	$\beta_3 TRV_{[t-1,t-5]}$	$\beta_4 TRV_{[t-1,t-22]}$
`Q-HAR-C-J`	$(\beta_1 + \beta_2 RQ^{1/2}_{t-1}) RV_{t-1} + \beta_3 J_{t-1}$	$\beta_4 TRV_{[t-1,t-5]} + \beta_5 J_{[t-1,t-5]}$	$\beta_6 TRV_{[t-1,t-22]} + \beta_7 J_{[t-1,t-22]}$
`QS-HAR-RV`	$(\beta_1 + \beta_2 RQ^{1/2}_{t-1}) RV^+_{t-1} + (\beta_3 + \beta_4 RQ^{1/2}_{t-1}) RV^-_{t-1}$	$\beta_5 RV_{[t-1,t-5]}$	$\beta_6 RV_{[t-1,t-22]}$
`QS-HAR-C`	$(\beta_1 + \beta_2 RQ^{1/2}_{t-1}) RV^+_{t-1} + (\beta_3 + \beta_4 RQ^{1/2}_{t-1}) RV^-_{t-1}$	$\beta_5 TRV_{[t-1,t-5]}$	$\beta_6 TRV_{[t-1,t-22]}$
`QS-HAR-C-J`	$(\beta_1 + \beta_2 RQ^{1/2}_{t-1}) RV^+_{t-1} + (\beta_3 + \beta_4 RQ^{1/2}_{t-1}) RV^-_{t-1} + \beta_5 J_{t-1}$	$\beta_6 TRV_{[t-1,t-5]} + \beta_7 J_{[t-1,t-5]}$	$\beta_8 TRV_{[t-1,t-22]} + \beta_9 J_{[t-1,t-22]}$

Each respective HAR model is given by: $\alpha + D_{t-1} + W_{t-1} + M_{t-1} + \varepsilon_t$. The forecasts are: $F_t = \alpha + D_{t-1} + W_{t-1} + M_{t-1}$, except for the `HAR-RVlog` that is given by $F_t = \exp\left\{\alpha + D_{t-1} + W_{t-1} + M_{t-1} + \frac{\sigma^2_\varepsilon}{2}\right\}$, where $\sigma^2_\varepsilon = \sum_{i=1}^{n} \frac{(\varepsilon_{t-i} - \bar{\varepsilon})^2}{n-1}$, which gives the variance of all past residuals (ε). σ^2_u is the conditional variance of the errors, ε.
The daily, weekly, and monthly components of volatility are represented by D_{t-1}, W_{t-1}, and M_{t-1}, respectively. Read the full explanation for what each component does in section 4.2.3.

Estimation and application

HAR models are tested with functions that I made, using the `lm` function of the "stats" R (R Core Team, 2020) package. The estimator for the linear regression is ordinary least squares (OLS)[23].

4.2.4 Univariate GARCH models

Nested GARCH models

In the Autoregressive Conditional Heteroskedasticity (ARCH) model (Engle, 1982), the conditional *return* process with zero-mean, r_t, is given by

$$r_t = \varepsilon_t = \sigma_t z_t, \;\; z_t \sim \mathcal{N}(0,1) \text{ and i.i.d.}, \tag{19}$$

where z_t is a Gaussian noise variable, and σ_t is the conditional volatility process that is itself defined by

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 . \tag{20}$$

In the simplest case ($q = 1$), the magnitude of innovations at t (ε_t^2) depends on the previous day's innovation (ε_{t-1}^2). This allows for heteroskedastic and autocorrelated errors in an uncorrelated return process. Bollerslev (1986) adds persistence to the volatility process by including the previous day's (modelled) variance, σ_{t-1}^2,

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 . \tag{21}$$

When $p = q = 1$, the model is very interpretable: the effect of ε_t on the future conditional volatility decays at a geometric rate of β when $0 < \beta < 1$; if the value of ε_{t-1}^2 is large, the conditional volatility increases, and if ε_{t-1}^2 is close enough to 0, the conditional volatility decreases.

Although there are a lot of extensions on this GARCH model, they often only differ in how the conditional volatility is affected by the error's *sign* and *magnitude*[24]. Each model does it differently, but the underlying structure tends to be similar, and that makes it possible to nest most popular GARCH models in the family GARCH (fGARCH) of Hentschel (1995):

[23] Clements and Preve (2019) show alternative ways of estimating HAR models (using weighted least squares with different weights, for example). I tested this approach and the results were consistently worse – the mean squared error and quasi-likelihood of every model got worse by roughly 8%.

[24] My explanation of the effects of volatility are based on the news impact curve (NIC), which is explained in more detail in the original fGARCH paper (Hentschel, 1995).

$$\frac{\sigma_t^\lambda - 1}{\lambda} = \omega + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}^\lambda (|z_{t-i} - \eta_{2i}| - \eta_{1i}(z_{t-i} - \eta_{2i}))^\delta + \sum_{i=1}^{p} \beta_i \frac{\sigma_{t-i}^\lambda - 1}{\lambda}, \quad (22)$$

where there is a Box-Cox transformation (Box and Cox, 1964) on the conditional volatility. The models I test from this family are in Table 6. All those models have $p = q = 1$ and zero-mean conditional return processes because those are variables that, empirically, do not meaningfully affect prediction (please refer to Hansen (2005) and the Literature Review chapter). The other four parameters of the models change how the conditional volatility responds to the residual of the return process. The first two parameters, δ and λ, make the conditional volatility increase or decrease depending on the standardized error's (z_t) *magnitude*. More specifically, as errors move away from 0, when $\delta = \lambda = 1$ (modelling the standard deviation) the variance increase linearly; when $\delta = \lambda = 2$ (modelling the variance) the variance increases exponentially with base 2; and so on.

The third parameter, η_1, creates asymmetry in response to the standardized error's *sign*. When η_1 is positive (which is typical in stock market data), the conditional volatility increases at a faster pace as the error becomes more negative, and increases at a slower pace as the error becomes more positive. In other words, when η_1 is positive, the response of conditional volatility to the values of z_{t-1} slopes up, and the response for positive values of z_{t-1} slopes down by the same amount (the shape dictated by δ and λ is unaltered). Because of the change in slope, η_1 has a larger effect on values that are farther away from 0.

It is, however, possible that the effect of η_1 is not centred on $z_{t-1} = 0$ due to the fourth parameter, η_2. In models where this parameter is free, the conditional volatility decreases the most when $z_{t-1} = \eta_2$, and the whole response to values of z_{t-1} shifts horizontally by -η_2 (for example, the effect on the conditional volatility of $z_t^2 = 1$ when $\eta_2 = 0$ is the same as the effect of $z_t^2 = 1.75$ when $\eta_2 = 0.75$). Because most of the probability density of returns is close to 0, small values of z_{t-1} are most responsible for the effect of η_2, which tends to cause (or correct) the bias in the volatility model. The variations on GARCH models mix and match all the characteristics imbued by the four parameters by either setting the parameters free or fixed.

The Exponential GARCH of Nelson (1991) is also a nested model, but the interpretation is not as straightforward. The `E-GARCH(p,q)` is defined as,

$$\log \sigma_t^2 = \omega + \sum_{i=1}^{q} (\alpha_i z_{t-i} + \gamma_i (|z_{t-i}| - \mathbb{E}|z_{t-i}|)) + \sum_{i=1}^{p} \beta_i \log \sigma_{t-i}^2, \quad (23)$$

where $\mathbb{E}|z_{t-i}| = 1$ when we assume that $z_t \sim \mathcal{N}(0,1)$. The point at which volatility lowers the most depends on the values of α_i and γ_i. Here, α_i induces asymmetry in an identical

way to -η_1. The effect of γ is also similar to the effect of η_1, except[25] that the change in slope is symmetrical: when γ is positive, the response of z_{t-i} to the conditional volatility slopes up by the same magnitude for $z_t < 0$ and $z_t > 0$.

In practice, all the nested models give similar results, with the main differences originating because of extreme returns and how the leverage effect is modelled.

Non-nested GARCH models

I test two GARCH-type models that don't fit in this family, although they are assumed to follow the same return process. The first one is the Component GARCH of Engle and Lee (1999), `C-GARCH`(p, q),

$$\sigma_t^2 = \widetilde{\omega}_t + \sum_{i=1}^{q} \alpha_i(\varepsilon_{t-i}^2 - \widetilde{\omega}_{t-i}) + \sum_{i=1}^{p} \beta_i(\sigma_{t-i}^2 - \widetilde{\omega}_{t-i}), \tag{24}$$

$$\widetilde{\omega}_t = \omega + \rho\widetilde{\omega}_{t-1} + \phi(\varepsilon_{t-1}^2 - \sigma_{t-1}^2), \tag{25}$$

where $\widetilde{\omega}_t$ is the long-term component of volatility. This model can be represented in an alternative way to aid understanding:

$$\sigma_t^2 = \widetilde{\omega}_t + s_t, \tag{26}$$

$$\widetilde{\omega}_t = \omega + \rho\widetilde{\omega}_{t-1} + \phi(\varepsilon_{t-1}^2 - \sigma_{t-1}^2), \tag{27}$$

$$s_t = (\alpha + \beta)s_{t-1} + \alpha(\varepsilon_{t-1}^2 - \sigma_{t-1}^2), \tag{28}$$

with s_t being the short-term component of volatility because it is assumed that $0 < (\alpha+\beta) < \rho < 1$. This model is particularly useful when there are extremely large volatility shocks that have a short-lived influence on future volatility, such as the 1987 stock market crash (Schwert, 1990).

The other model is the Realized GARCH, `R-GARCH`(p, q), of Hansen, Huang, and Shek (2012) that uses realised volatility with the (log-linear) specification,

$$\log \sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \log \mathrm{RVol}_{t-i} + \sum_{i=1}^{p} \beta_i \log \sigma_{t-i}^2, \tag{29}$$

$$\log \mathrm{RVol}_t = \xi + \delta \log \sigma_t^2 + \tau(z_t) + u_t, u_t \sim N(0, \sigma_u^2), \tag{30}$$

$$\tau(z_t) = \eta_1 z_t + \eta_2(z_t^2 - 1), \tag{31}$$

where RVol_t is a measure of realised volatility; ξ is the constant term of that process; $\tau(z_t)$

[25] There is another exception: a change in γ also causes the whole response of conditional volatility to move up or down, but ω ends up adjusting for that.

Table 5: Nested GARCH models (fGARCH)

Model name	Author(s)	Model notation	p	q	λ	δ	η_1	η_2
ARCH	Engle (1982)	`ARCH(q)`	0	1	2	2	0	0
Generalized ARCH (GARCH)	Bollerslev (1986)	`GARCH(p, q)`	1	1	2	2	0	0
Integrated in variance GARCH	Engle and Bollerslev (1986)	`I-GARCH(p, q)`	1	1	2	2	0	0
Glosten, Jagannahan, and Runkle GARCH	Glosten et al. (1993)	`GJR-GARCH(p, q)`	1	1	2	2	Free	0
Nonlinear asymmetric ARCH	Engle and Ng (1993)	`NA-ARCH(p, q)`	1	1	2	2	0	Free
Absolute Value GARCH	Taylor (2008, Ch. 3); Schwert (1989)	`AV-GARCH(p, q)`	1	1	1	1	Free	Free
Threshold GARCH	Zakoian (1994)	`T-GARCH(p, q)`	1	1	1	1	$\lvert\eta_1\rvert \geqslant 1$	0
Asymmetric Power ARCH	Ding, Granger, and Engle (1993)	`AP-ARCH(p, q)`	1	1	Free	$\delta = \lambda$	$\lvert\eta_1\rvert \geqslant 1$	0
Nonlinear ARCH	Higgins and Bera (1992)	`N-ARCH(p, q)`	1	1	Free	$\delta = \lambda$	0	0
Full Family GARCH	Hentschel (1995)	`F-GARCH(p, q)`	1	1	Free	$\delta = \lambda$	Free	Free

These parameters are from the equation (22) which I put here for easy reference:
$\frac{\sigma_t^\lambda - 1}{\lambda} = \omega + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}^\lambda (|z_{t-i} - \eta_{2,i}| - \eta_{1,i}(z_{t-i} - \eta_{2,i}))^\delta + \sum_{i=1}^{p} \beta_i \frac{\sigma_{t-i}^\lambda - 1}{\lambda}$.
The meaning of each parameter is explained after equation (22) in the body of the text.

is the component that models the leverage effect; σ_u^2 is the variance of the residuals of the realised volatility process, u_t, which is jointly estimated with every other parameter; and $\tau(z_t)$ is the leverage function. Like in the fGARCH specification given by equation (22), η_1 creates an asymmetry in the response of negative or positive standardized residuals, z_t; η_2 has a similar effect to γ_i in the `E-GARCH` because the effect is symmetric, but here it affects z_t^2 instead of $|z_t|$. The main feature of this model is that the estimation of the realised volatility is affected both by the log variance ($\log \sigma_t^2$) and residuals of the return process through $\tau(z_t)$. The model also produces estimates for RV_t and r_t^2, so they are shown separately in the form of $\texttt{R-GARCH}_t^{RV}(1,1)$ and $\texttt{R-GARCH}_t^{r^2}(1,1)$, respectively.

Estimation and implementation of the models

All the GARCH models are implemented using the Rugarch package (Ghalanos, 2014) for the R programming language (R Core Team, 2020), and always assuming that the errors are normally distributed. For prediction, this package is the most complete according to Hill and McCullough (2019). The estimation of the parameters is done with the typically used quasi-maximum likelihood (White, 1982).

4.3 Model combinations

As I argued in section 4.2, some information in the data may only be exploited when conditioning on other variables (e.g., volatility jumps may improve performance only when the model corrects for measurement errors). The forecasts of individual models are highly collinear, much more so than most realised measures, so model combinations are a blunt tool to explore these masked relationships. As I also argued before, the approach of trying to build a "super model" has not yielded many tangible benefits – GARCH models see little-to-no improvements in performance when they have more parameters or different conditional distributions, and the best HAR models, such as the `Q-HAR` of Bollerslev et al. (2016), seldom benefit from using more proxies for volatility measures[26].

This means that combining models remains one of the few methods that could plausibly improve forecast performance and, at the same time, reduce the uncertainty in the model selection process. To achieve this goal, the combination methods I use are all meant to reduce the variance of the forecasts because the high collinearity between forecasts adds difficulties in estimating the independent effect of one model relative to others.

Every model combination that needs to be estimated on the performance of individual models either uses the mean squared error (MSE) loss function or the quasi-likelihood

[26] I still test HAR models with many volatility proxies because there has been no sufficiently thorough empirical test, as I argued in section 4.2.3.

Table 6: Number of model combinations

Model family	Number of models
Grouping	127
Trimming	42
Regularization	6
Equal Weight	1

(QLIKE) loss function. I want to test this to see how these methodologies optimize for different loss functions, and it also serves as a small robustness test.

Any negative forecast equal to or smaller than zero is set to half the previous day's forecast.

4.3.1 Equal weighting and grouping

Let V_t contain all the individual models' forecasts for the day t, $F_{t,i}$:

$$V_t \equiv \{F_{t,i}\} \quad \text{for} \quad i = 1, 2, ..., n. \tag{32}$$

The forecast for the equal weighting combination, $\texttt{EqualWeight}_t$, is a simple average of all individual forecasts:

$$\texttt{EqualWeight}_t \equiv (\#V_t)^{-1} \sum_{F_{t,i} \in V_t} F_{t,i}\,. \tag{33}$$

The disadvantage of this weighting scheme is that it is too naïve – it assigns a $\frac{1}{n}$ weight to the only implied volatility model and assigns the same $\frac{1}{n}$ weight for each of the 11 (very correlated) GARCH models. In practice, this means that the weighting scheme overrepresents model categories that have a higher number of individual models.

As an attempt to fix this, I use a slightly more elaborate weighting scheme to adjust the weights based on the *category* the model belongs to implied volatility (IV), naïve models with squared returns (Naïve^{r}) and realised variance (Naïve^{RV}), ARMA models (ARMA^{r} or $\texttt{ARMA}^{RV}$), GARCH models (GARCH), and HAR models (HAR). I choose to separate them this way because the literature has shown that the model structure and the volatility proxy used are two important characteristics that meaningfully distinguish these models.

The forecast for each category is simply an equally weighted average of the models in that category. For example, the forecast for what I call a "grouping" model containing only

ARMA (RV) models is given by

$$\texttt{Grouping}(\{\text{ARMA}^{\text{RV}}\}) \equiv \frac{1}{3}\left(\texttt{ARMA}^{\text{RV}}(0,1)_t + \texttt{ARMA}^{\text{RV}}(1,0)_t + \texttt{ARMA}^{\text{RV}}(1,1)_t\right).$$

I can combine more than one model category. For example, I can combine the category of ARMA models using realised variance, ARMA^{RV}, and the category of GARCH models, GARCH. When combining multiple model categories, each category's forecast is assigned a weight inversely proportional to the number of categories being used (i.e., $\frac{1}{2}$ for 2 categories, $\frac{1}{3}$ for 3 categories, and so on). In the example above, the combination assigns 50% of the weight to at the forecasts of ARMA^{RV} (which is split among its 3 models), and 50% to GARCH (which is split among its 11 models):

$$\begin{aligned}\texttt{Grouping}(\{\text{ARMA}^{\text{RV}}, \text{GARCH}\}) \equiv &\frac{0.5}{3}(\texttt{ARMA(1,0)}^{\text{RV}}_t + \texttt{ARMA(0,1)}^{\text{RV}}_t + \texttt{ARMA(1,1)}^{\text{RV}}_t) \\ &+ \frac{0.5}{11}(\texttt{ARCH(1)} + \texttt{GARCH(1, 1)} + ... + \texttt{R-GARCH(1, 1)}).\end{aligned}$$

More generally, let the set B contain all the categories that are going to be included in a given forecast, and the forecasts for those grouping models be given by

$$\texttt{Grouping}(B) \equiv \sum_{V_{i,t} \in B} \frac{(\#V_{i,t})^{-1}}{\#B} \sum_{F_{i,t} \in V_{i,t}} F_{i,t}, \tag{34}$$

where $V_{i,t}$ contains all the forecasts of one of the 7 model categories, indexed by $i = 1, 2, ..., 7$ in no particular order. I test all the possible models that are possible to make with these categories, so there are $\binom{7}{\#B}$ possible combinations for each unique set B that can be made with the 7 categories (127 models in total).

Estimation and application

This model does not require estimation. It is implemented using the R programming language (R Core Team, 2020) with functions I created.

4.3.2 Trimming

Trimming[27] is the practice of removing a portion of models that have underperformed in the past. In the literature, it is common to remove models whose (cumulative) loss is below a certain percentile (e.g., for the forecast combination at time t, only include model i if $\text{MSE}_{i,t-1}$ is below the 80^{th} percentile of all $\text{MSE}_{i,t-1}$).

[27] "Trimming" is often used differently in the literature; read the footnote 0 on page 0 for more details.

The problem with this approach is that it is hard to compare results across studies – a researcher only comparing GARCH and HAR models might find that it is better to exclude 10% of models, while another researcher that is also testing naïve models might need to exclude 50% of models to get the best performance. To try to have a more stable measure, I exclude models that exceed a threshold relative to the previous day's minimum (cumulative) loss function (e.g., exclude the i^{th} model if $MSE_{i,t-1}$ is above $1.25 \times \min[MSE_{j,t-1}]$), and the forecasts of the models that are not excluded receive a weight inversely proportional ($\frac{1}{n}$) to the number of remaining models (n).

I denote this threshold by k, and I test the values of $1, 1.1, ..., 3$ for it. When $k = 1$, only the models with the lowest loss are selected; when $k = 3$, very few models are excluded in my data set (on average, less than 1 forecast is excluded in each day). Formally:

$$\texttt{Trimming}(k)_t \equiv n^{-1} \sum_{i=1}^{n} F_{i,t} | L_{i,t-1} <= (k \times \min L_{\cdot,t-1}), \tag{35}$$

where $L_{i,t-1}$ is the loss function used (MSE or QLIKE) for the previous day, and $F_{i,t}$ represents the forecast given by some individual model i.

Estimation and application

This procedure is also implemented in the R programming language (R Core Team, 2020). As I described in section 3, the individual models make out-of-sample predictions for the first 2 years, so the trimming model combinations are trained on those 499 observations. This ensures that the trimming is not excluding models based on overfit in-sample data. This process is repeated for every observation with an expanding window.

4.3.3 Regularization: Ridge, Lasso, and Elastic Net regression

With the typically used ordinary least squares (OLS) estimator, the coefficients ($\hat{\beta}_t^{OLS}$) are found by minimizing the residual sum of squares (RSS), here with the notation adapted to time series:

$$\begin{aligned} \hat{\beta}_t^{OLS} &= \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=0}^{N-1} (y_{t-i} - \beta_0 - \sum_{j=1}^{p} \beta_j x_{t-i,j})^2 \right\} \\ &= \underset{\beta}{\operatorname{argmin}} \{RSS\}, \end{aligned} \tag{36}$$

where y_{t-i} is the response variable (the realised variance in my case), β_0 is the constant term, β_j is the coefficient of a predictor (the forecasts of some individual model, in my case), and $x_{t-i,j}$ are the specific observations for those predictors.

Ordinary least squares gives unbiased coefficient estimates, even when heteroskedasticity and multicollinearity are present (Wooldridge, 2016, p. 82-86). But when we break those assumptions and/or when the sample size is relatively small, the estimator gives estimates with high variance – a small change in one observation can cause large changes in the estimated parameters (Hastie, Tibshirani, and Friedman, 2009, p. 219-231). When that variance is primarily driven by the noise in the data, the parameter estimates cannot be generalized to data outside the sample. By applying some form of regularization, we are shrinking the parameters towards zero, so we are effectively assuming that there is some reversion to the mean – on average, the true parameter's value is less extreme than the estimated parameter.

The ridge regression (Hoerl and Kennard, 1970; Hastie et al., 2009, p. 61-68) applies shrinkage to the squared values of coefficients obtained by minimizing the residual sum of squares (RSS):

$$\hat{\beta}_t^{\text{ridge}} = \underset{\beta}{\text{argmin}} \left\{ \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{37}$$

where the penalty to the *squared* value of the coefficients is given by λ (ℓ_2 penalty). The Least absolute shrinkage operator (Lasso) (Tibshirani, 1996; Hastie et al., 2009, p. 68-69) is similar, but it penalizes the *absolute* value of the coefficients (ℓ_1 penalty),

$$\hat{\beta}_t^{\text{lasso}} = \underset{\beta}{\text{argmin}} \left\{ \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{38}$$

Lastly, the Elastic Net[28] (Zou and Hastie, 2005; Hastie et al., 2009, p. 661-664) combines both these penalties by giving a weight of α to the ℓ_1 penalty, and a weight of $1 - \alpha$ to the ℓ_2 penalty[29]:

$$\hat{\beta}_t^{\text{enet}} = \underset{\beta}{\text{argmin}} \left\{ \text{RSS} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right) \right\}. \tag{39}$$

All the methods' forecasts are simply given by the sum of the parameters estimated with their respective methodologies, so the formulas are the same (with the only difference that

[28] I use the "naïve Elastic Net", as Zou and Hastie (2005) call it, since that is the version of the model that is used in applications. The naïve Elastic Net is more biased and less variable than the other version.

[29] We can ignore the $\frac{1}{2}$ that is multiplied by $1 - \alpha$ because that does not affect the calculation of the weights.

p, the number of parameters, might be different when $\alpha > 0$):

$$\texttt{ElasticNet}(\lambda, \alpha)_t = \beta_0 + \sum_i^p \beta_i F_{t-1,i}\,; \tag{40}$$

$$\texttt{ElasticNet}(\lambda, 0)_t = \texttt{Ridge}(\lambda)_t = \beta_0 + \sum_i^p \beta_i F_{t-1,i}\,; \tag{41}$$

$$\texttt{ElasticNet}(\lambda, 1)_t = \texttt{Lasso}(\lambda)_t = \beta_0 + \sum_i^p \beta_i F_{t-1,i}\,; \tag{42}$$

where $F_{t-1,i}$ is the forecast of some individual model. In practice, the λ coefficient varies from day to day because of the estimation procedure (read more about it in the Estimation and application subsection below).

For the ridge regression, the penalty is given to β_j^2, so larger values of coefficients get disproportionately penalized, whereas values close to zero receive a small penalty. In practice, this means that the coefficients tend to become similar to each other and/or shrunk very closely towards 0 (while never becoming exactly 0). This is advantageous when all or almost all the predictors are informative, and when there is high multicollinearity. The latter case is especially relevant for this study because the OLS estimator tends to make some of the coefficients extremely positive or negative, even if they have very similar information. When using ridge regression, they tend to become all positive and small if the shrinkage parameter (λ) is sufficiently large.

Because the Lasso applies a penalty to the absolute values of the coefficients ($|\beta_j|$), it more quickly shrinks some of the coefficients towards zero, and they can become exactly zero. This is optimal when few predictors contain relevant information for forecasting (i.e., the solution is sparse). In the presence of multicollinearity, however, it tends to pick the collinear predictor that happened to be better than others and shrink all the relative underperformers to 0. This is seldom desirable because multicollinear predictors often have identical information, so they should be averaged to some extent.

To get to a compromise between both approaches, the Elastic Net applies both types of regularization – it first applies the ℓ_1 regularization (removes some of the unimportant predictors), and then applies the ℓ_2 regularization (which gives all the advantages I previously mentioned).

These three techniques are equivalent to a wide set of methodologies used in the literature. Hastie et al. (2009, p. 69-79) show that the ridge regression is identical to a continuous version of principal components regression (PCR) and partial least squares regression (PLSR), and Stock and Watson (2012) show that this type of shrinkage (penalty to coefficients

estimated by OLS, which includes some factor models) yields identical results, independently of the specific form in which they are applied. Hastie et al (2009, p. 69-79) also show that the Lasso is similar to a continuous version of subset selection methods and least angle regression (LAR), and the elastic net has a similar effect to the bridge regression (the bridge applies shrinkage to $|\beta_j|^p$, so it gives similar values to the elastic net when $1 < p < 2$, except that these values of p do not eliminate parameters, unlike the Elastic Net when $\alpha > 0$). Finally, these regularization techniques give solutions that can be obtained with a Bayesian approach to model averaging (by applying data-driven Gaussian or Laplacian priors for the Ridge and Lasso regressions, respectively), which Stock and Watson (2012) also tested.

Estimation and application

I use the R package "glmnet" (Friedman, Hastie, and Tibshirani, 2010) for the regression, and my functions for the time series leave-one-out cross-validation (which I call tsCV) and grid search procedures.

I need to use tsCV because the typical cross-validation procedure is only a good approximation of out-of-sample performance if there is no variable selection based on performance, and the observations in the observations in the sample are independent and identically distributed[30] (Hastie et al, 2009, p. 241-249; Arlot, Celisse, et al, 2010). There are ways of adapting to that dependence structure, but Bergmeir and Benítez (2012) find that they are inferior to the so-called "blocked cross-validation", which is often referred to as "time series cross-validation". This procedure is similar to the "last block evaluation" (or "hold out sample") procedure, where there are separate training and testing data sets, but here the model is retrained every k observations. I use tsCV with $k = 1$ given by

$$\text{tsCV}^{\text{MSE}} \equiv n^{-1} \sum_{t=w}^{n} (f(\mathbf{X}_{t-1}, \mathbf{Y}_{t-1})_t - y_t)^2, \tag{43}$$

$$\text{tsCV}^{\text{QLIKE}} \equiv n^{-1} \sum_{t=w}^{n} \left(\frac{y_t}{f(\mathbf{X}_{t-1}, \mathbf{Y}_{t-1})_t} - \log \left(\frac{y_t}{f(\mathbf{X}_{t-1}, \mathbf{Y}_{t-1})_t} \right) - 1 \right), \tag{44}$$

where $f(X_{t-1}, Y_{t-1})_t$ is the forecast for day t, which uses data up to $t-1$ (i.e., $Y_{t-1} \equiv \{y_t\}$ for $t = 1, 2, ..., t-1$; and $X_{t-1} \equiv \{x_{t,s}\}$ for $t = 1, 2, ..., t-1$ and $s = 1, 2, ..., p$). Crucially, w is the day where tsCV starts, and n the day where it ends. Assuming that $t = 1$ corresponds to 01-01-2005 (the day where individual models first make their out-of-sample predictions), it is always the case that $w = 249$. Initially, $n = 500$ but it increases by 1 every time the model makes a step-ahead forecast.

[30] This is not strictly true for time series, as Bergmeir, Hyndman, and Koo (2018) have shown, but they find that to use cross-validation we need to use purely (non-linear) autoregressive models with uncorrelated errors, which is not an assumption that holds in my case.

To find the optimal values of λ and α, I have to repeat the cross-validation procedure multiple times and then pick the optimal combination[31]. If there are ties in tsCV, I randomly choose one set of hyperparameters from the ones that tied. Because there are only two hyperparameters, I use grid search, which just involves defining a (discrete) range of values for those hyperparameters and then testing all possible combinations of those values.

In my case, α can take the values of $0, 0.05, ..., 1$, and λ the values of $\exp(w)$ where $w = -28, -27.8, ..., -12$. I pick these values because they were very wide, and widening them further did not affect the results.

This whole procedure is done every day, so the values of λ and α for the Elastic Net, and the values of λ for the Ridge and Lasso can also change every day. There is also another detail related to how the variables are standardized[32]. The shrinkage works on the values of the coefficients, but if these are in different scales, most of the penalty goes for the predictor with the largest scale. This is not the case with my study, so I do not standardize the predictors. The standardization would still lead to differences in the results because not all variables have the same variance – relative to high-variance predictors, low-variance predictors are *less* penalized after standardizing. But I do not want this to happen. I prefer that, for example, a naïve model with a long-term average suffers more shrinkage because their (small) fluctuations are likely noisy, so their performance is unlikely to be good. At most, those less variable models might help with bias-correction in naïve combination schemes.

4.4 Significance test: Model confidence set

I represent the set of all models (which includes individual models or combinations) by $\mathcal{M}^0$, where models are indexed by $i = 1, 2, ..., m_0$. Let the error for the i^{th} model at time t be given by $L_{i,t}$ (which can either be the standard error, $SE_{i,t}$, or the quasi-likelihood error, $QLE_{i,t}$, as described in section 4.1). Let $d_{ij,t}$ be the difference in loss function between the i^{th} and j^{th} models at time t:

$$d_{ij,t} \equiv L_{i,t} - L_{j,t}, \text{ for all } i, j \in \mathcal{M}^0 . \tag{45}$$

The MCS procedure tests the null hypothesis that the expected difference in the mean loss function between models is 0,

$$H_{0,\mathcal{M}} : \mathbb{E}(d_{ij}) = 0 \text{ for all } i, j \in \mathcal{M} , \tag{46}$$

[31] To prevent overfitting, which happens in all forms of model selection to some extent Cawley and Talbot (2010), it is common to apply the one-standard-deviation heuristic, defined as picking the value λ that gives the most parsimonious model with an error (at most) one standard error away from the best result (Hastie et al., 2009, p. 61). I planned to do this, but this rule only picked an intercept-only model, so I also look at the performance of the regularization methods by looking at how the out-of-sample performance varied with all values of lambda.

[32] This can be applied in time series so long as the data that is used for standardization for a model at time t only includes data up to $t - 1$.

where $\mathcal{M} \subset \mathcal{M}^0$. The MCS procedure sequentially tests this hypothesis of equal predictive ability on pairs of models and eliminates the models from $\mathcal{M}$ that are found to be significantly inferior. The objective of the test is to find the set of superior models, $\mathcal{M}^*$, defined by

$$\mathcal{M}^* \equiv \{i \in \mathcal{M}^0 : \mu_{ij} < 0 \quad \text{for all} \ \ j \in \mathcal{M}^0\}. \tag{47}$$

In practice, the set of best models is unknown, so we try to estimate what models are in that set for a given level of confidence (α) such that the models in $\mathcal{M}^*$ are in $\widehat{\mathcal{M}}^*_{1-\alpha}$ at least $(1-\alpha)$% of the samples (i.e., $P(\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha} \geqslant 1-\alpha)$).

Let $\bar{d}_{ij}$ be the average of the differences in loss function over the n periods for the i^{th} and j^{th} models,

$$\bar{d}_{ij} \equiv n^{-1} \sum_{t=1}^{n} d_{ij,t}\,; \tag{48}$$

and let $\bar{d}_{i.}$ be the average of $\bar{d}_{ij}$ for the m models inside $\mathcal{M}$, which is the set of models that have not been excluded,

$$\bar{d}_{i.} \equiv m^{-1} \sum_{j \in \mathcal{M}} \bar{d}_{ij}\,; \tag{49}$$

the statistic that I use for the procedure is the mean of the (standardized) difference between the loss of the i^{th} relative to the j^{th} model inside $\mathcal{M}$:

$$t_{i.} \equiv \frac{\bar{d}_{i.}}{\sqrt{\widehat{var}(\bar{d}_{i.})}}, \tag{50}$$

where $\widehat{var}(\bar{d}_{i,j})$ is a bootstrapped estimate of $var(\bar{d}_{i,j})$. The excluded model is the one that has the highest loss relative to other models,

$$T_{max} \equiv \max_{i \in \mathcal{M}} t_{i.}. \tag{51}$$

With this, it is possible to give a detailed description of the MCS algorithm (Becker and Clements, 2008; Bernardi and Catania, 2018; Hansen, Lunde, and Nason, 2003; Hansen et al., 2011), which is the following:

1. Start with all the models in the MCS (i.e., $\mathcal{M} = \mathcal{M}_0$).

2. Calculate the T_{max} statistic as in equation (51). This statistic represents the model that got the worst (mean) loss when compared with all other models in $\mathcal{M}$.

3. Calculate the MCS p-value, $\hat{p}_i$. To do this, firstly resample the data with replacement B times (I use $B = 5000$) with a block bootstrap procedure. Secondly, calculate the T_{max} for each of the samples. Thirdly, to finally get the MCS p-value, calculate the per-

centage of times that $T_{max}^{(b)} > T_{max}$, where $T_{max}^{(b)}$ is the statistic for the b^{th} bootstrapped sample:

$$T_{max}^{(b)} \equiv \max_{i \in \mathcal{M}} \frac{\bar{d}_{i\cdot}^{(b)} - \bar{d}_{i\cdot}}{\sqrt{\text{var}\left(\bar{d}_{i\cdot}^{(b)} - \bar{d}_{i\cdot}\right)}}. \tag{52}$$

If T_{max} is very large, it becomes unlikely that $T_{max}^{(b)}$ is even larger because the $\bar{d}_{i\cdot}^{(b)}$ statistic is subtracted by $\bar{d}_{i\cdot}$. .

More succinctly, the MCS p-value is given by:

$$\hat{p}_i \equiv B^{-1} \sum_{b=1}^{B} 1(T_{max}^{(b)} > T_{max}), \tag{53}$$

$$1(A) \equiv \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{if } A \text{ is false.} \end{cases} \tag{54}$$

4. If $\hat{p}_i < \alpha$, remove the i^{th} model from $\mathcal{M}$. Else, if $\hat{p}_i \geqslant \alpha$, stop the procedure and set $\mathcal{M} = \widehat{\mathcal{M}}_{1-\alpha}^{*}$.

To compute T_{max}, the MCS procedure uses the performance measures of all models, so the results of the test can retain a general meaning without losing too much statistical power.

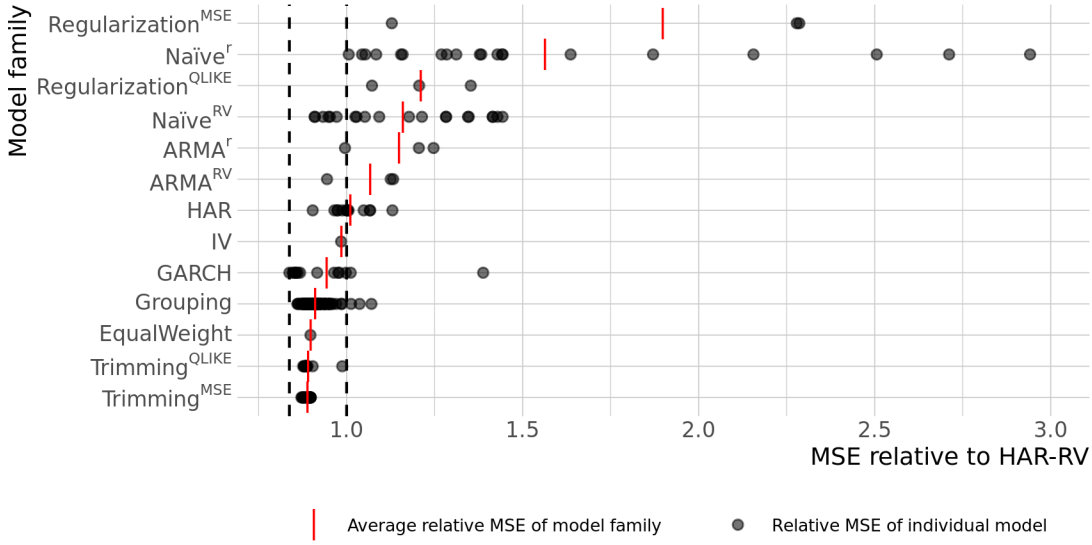
Estimation and application

This procedure is estimated with the R package "MCS" by Bernardi and Catania (2018). I include the 1004 observations from the out-of-sample period that has model combinations and individual models (though I exclude the duplicates, which only happens with the "grouping" model of implied volatility). The block bootstrap procedure uses 10000 simulations. To choose the block length, an AR(p) process is fit on all d_{ij} terms, and the number of blocks p is determined by the maximum number of significant parameters (as described by Bernardi and Catania (2018), which cite all the relevant literature), which resulted in a block length of 30. I set $\alpha = 0.2$; this is slightly higher than the typical 10% level chosen because I find the test to be too conservative (please refer to section 2.5), but using 10% does barely has any impact in the models that were excluded.

5 Empirical results

5.1 Summary

Figure 2: Out-of-sample MSE for individual models and model combinations (2007-2010)



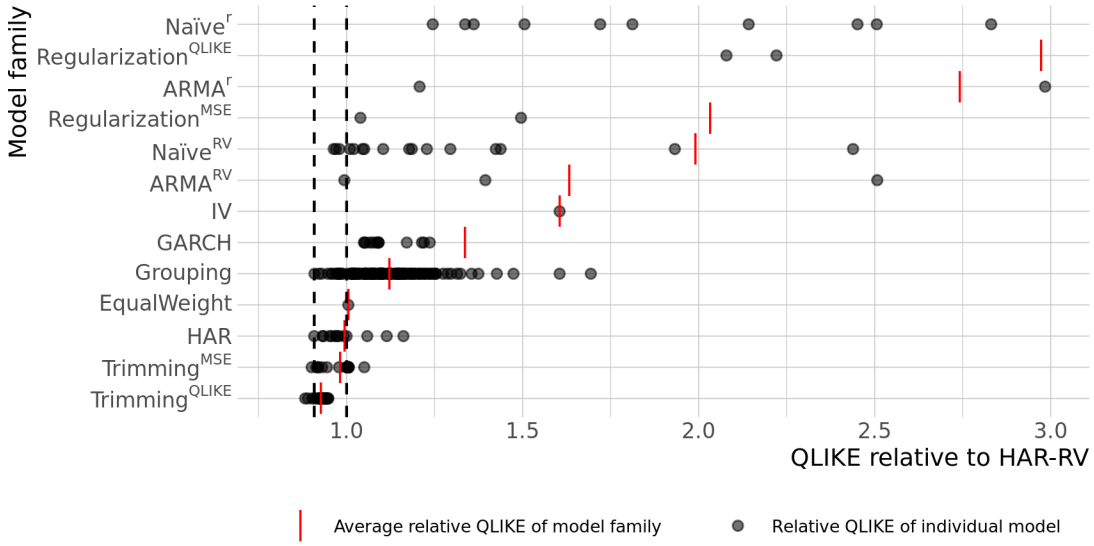
This figure compares the performance of all models I tested (black points), where each model category has a corresponding average of all those models (vertical red bars). This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH(1,1)` with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

All of my results are summarised in Figures 2 and 3, and the correlations between the model's forecasts are compared in Table 8 of Appendix B. For both the mean squared error (MSE) and the quasi-likelihood (QLIKE) loss functions, individual models are, on average, inferior model combinations (with exception of regularization methods). Combining models by excluding past underperformers ("trimming") is generally the best approach, but all simple model combination schemes lead to similar results: they usually perform better than the `HAR-RV` (represented by the rightmost vertical dotted line) but underperform the best individual model (which is represented by the leftmost vertical dotted line).

Out of the individual models, GARCH models appear to be the best alternative to practitioners who only have access to squared returns, since they fare similarly to HAR and ARMA-RV models. This is unusual in the literature – when compared to GARCH models, HAR models typically reduce the MSE by 20%-60% (please refer to the footnote on page 4).

The following characteristics are the only ones associated with slightly better perfor-

Figure 3: Out-of-sample QLIKE for individual models and model combinations (2007-2010)



This figure compares the performance of all models I tested (black points), where each model category has a corresponding average of all those models (vertical red bars). This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix □.

mance: allowing for the leverage effect, both with GARCH (e.g., $\texttt{E-GARCH}(1,1)$) and HAR (e.g., `S-HAR`) models; HAR models that have continuous and jump components, or that use the logarithm of realised variance (RVlog) instead of just the realised variance; and exponential smoothing models using realised volatility with high or intermediate persistence in the volatility process (i.e., $0.5 \leqslant \beta \leqslant 0.95$).

The Model Confidence Set (MCS) with an α level[33] set to 0.2 did not exclude any model from the set of superior models with the MSE, and it only excluded 3 models with the QLIKE: the random walk model with squared returns ($\texttt{RandWalk}^{\text{r}}$), and two exponential smoothing models ($\texttt{ExpSmooth}(0.3)^{\text{r}}$ and $\texttt{ExpSmooth}(0.4)^{\text{r}}$).

In the next section, I discuss in more detail the issues with this test. In light of those issues, and despite the lack of significant results, I expect that analyzing more subtle relationships will be informative for future research – if this were not the case, my results would probably not have replicated many of the patterns observed in the literature. That said, the analysis is still meant to be understood as primarily descriptive because my sample is relatively limited when compared to other studies in the literature. Additionally, although the

[33] It is typical to set $\alpha = 0.1$, but I believe the MCS test is too conservative, so I set a lower threshold for rejection. But there are barely any differences between using both levels in my case.

FTSE100 index is less studied than equity indexes from the United States of America, both of them will be correlated, so my results are not fully independent from the ones observed in the literature.

5.2 Individual models

5.2.1 Naïve models and implied volatility

The naïve model results in the 2007-2010 period (Figures 4 and 5) show similar tendencies to those observed in the literature: using squared returns is inferior to using realised volatility because squared returns are noisy (Andersen et al., 2005, 1999, 2003); some of the models with too much persistence (such as historical volatility and long-term moving averages) or too little persistence (such as the random walk model) tend to be substantially worse than other models (Poon and Granger, 2003); and this last tendency is more extreme with the QLIKE loss function – some naïve models have a QLIKE 2000 times larger than the QLIKE of the `HAR-RV` – because the QLIKE penalizes underpredictions more (e.g., Bollerslev et al. (2016) observes this tendency for the `AR(1)` model).

In my sample, I find a slight contradiction with the results of Poon and Granger (2005, 2003) because, while some naïve models do relatively better than GARCH models during the low volatility period of 2005-2006 (Appendix C), that relationship is driven by the worse GARCH performance during that period – the performance of naïve models remains roughly constant relative to the performance of HAR models.

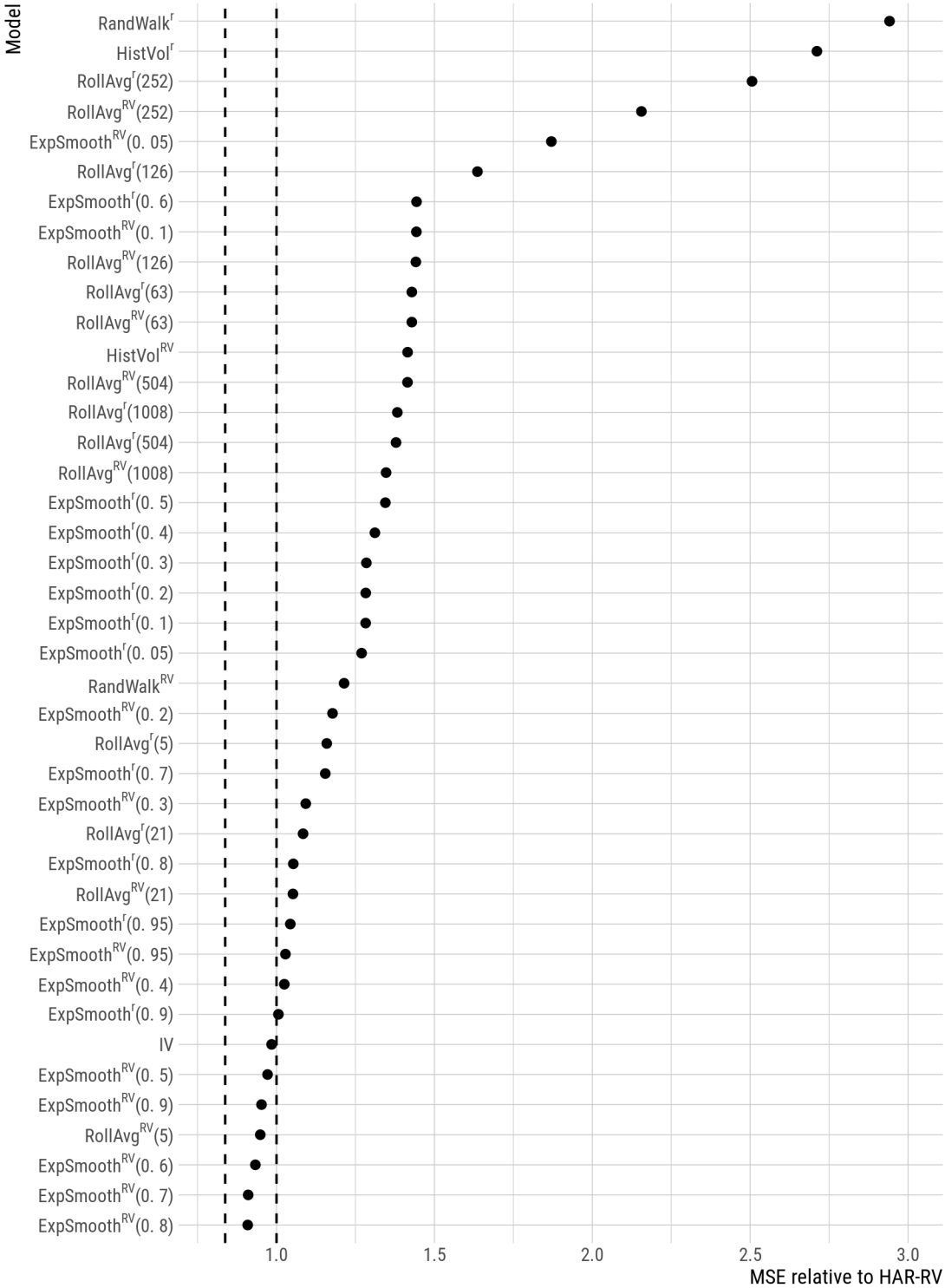
Implied volatility does slightly better than most naïve models, but it gets worse for the QLIKE loss function. This is slightly unexpected, as implied volatility has an upward bias (Chernov, 2007; Christensen and Prabhala, 1998), which should favour a loss function that values overpredictions more, as observed in Becker and Clements (2008). That said, the deterioration in performance is not large, and it is typical for implied volatility to perform slightly worse than statistical models for one-day-ahead forecasts.

Lastly, for exponential smoothing, both when parameters allow for high or medium persistence in the volatility process ($\beta > 0.5$ and $\beta \approx 0.5$, respectively), and when using the MSE or the QLIKE, the performance is on average very close to the `HAR-RV`.

5.2.2 HAR and ARMA with realised volatility

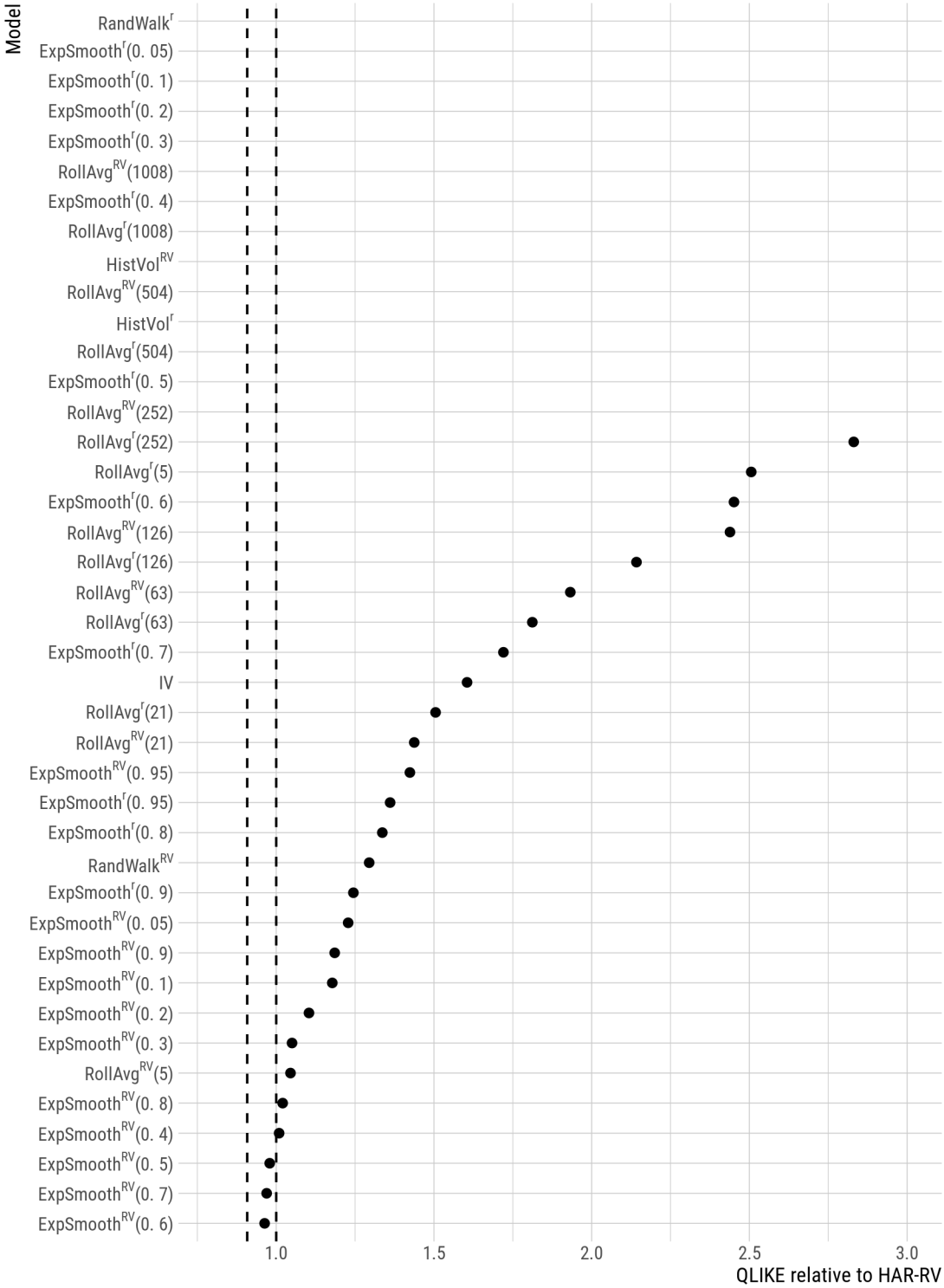
Both HAR and ARMA models have a much narrower range in their relative performances (Tables 6 and 5). For both categories, the simpler models (i.e., those with fewer parameters) do slightly and consistently worse, except for the `HAR-RVlog`, whose outperformance is mostly observed with the mean squared error – a pattern also observed in Buccheri and Corsi (2019). For the 2005-2006 period (Appendix C), the ordering of the models remains

Figure 4: Out-of-sample MSE for naïve models and implied volatility (2007-2010)



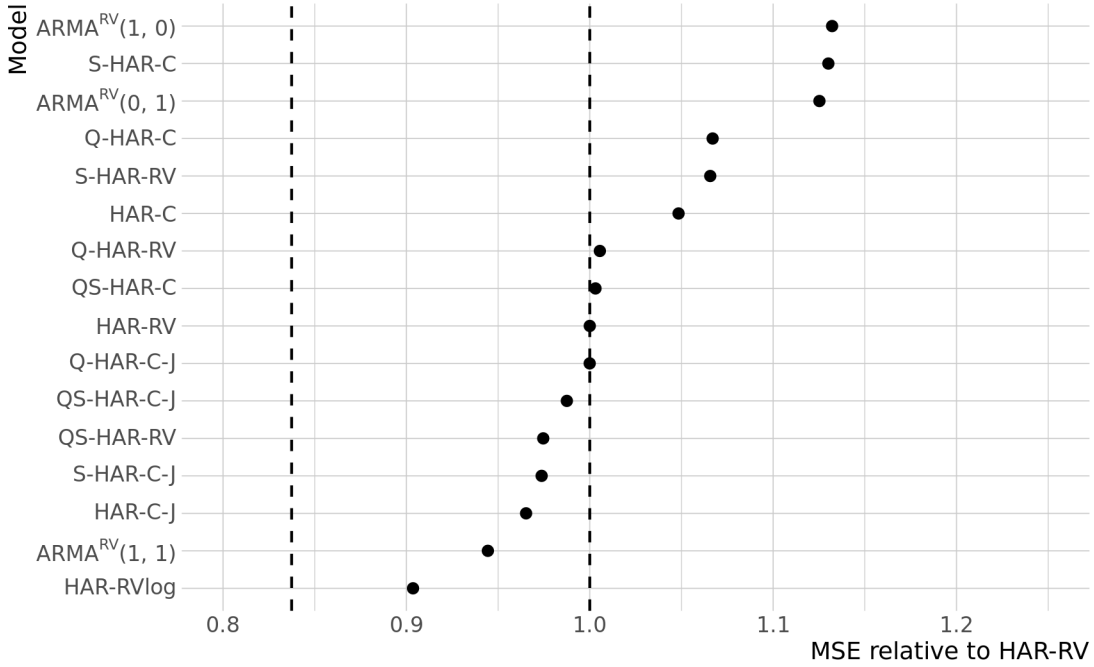
This figure compares the performance of all naïve models (4.2.1) I tested, ordered by their performance. This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH`$(1,1)$ with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

Figure 5: Out-of-sample QLIKE for naïve models and implied volatility (2007-2010)



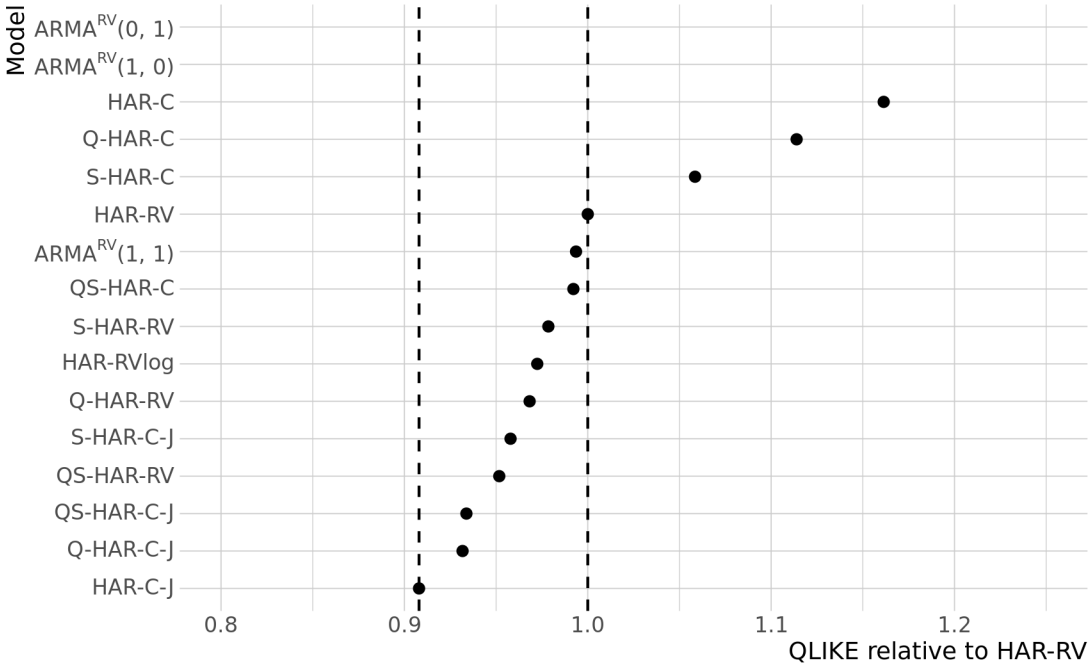
This figure compares the performance of all naïve models (4.2.1) I tested, ordered by their performance. This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix D.

Figure 6: Out-of-sample MSE for HAR and ARMA^{RV} models (2007-2010)



This figure compares the performance of all HAR models (section 4.2.3) and all ARMA models with realised variance (section 4.2.2) I tested, ordered by their performance. This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH`$(1,1)$ with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). These results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

Figure 7: Out-of-sample QLIKE for HAR and ARMA^{RV} models (2007-2010)



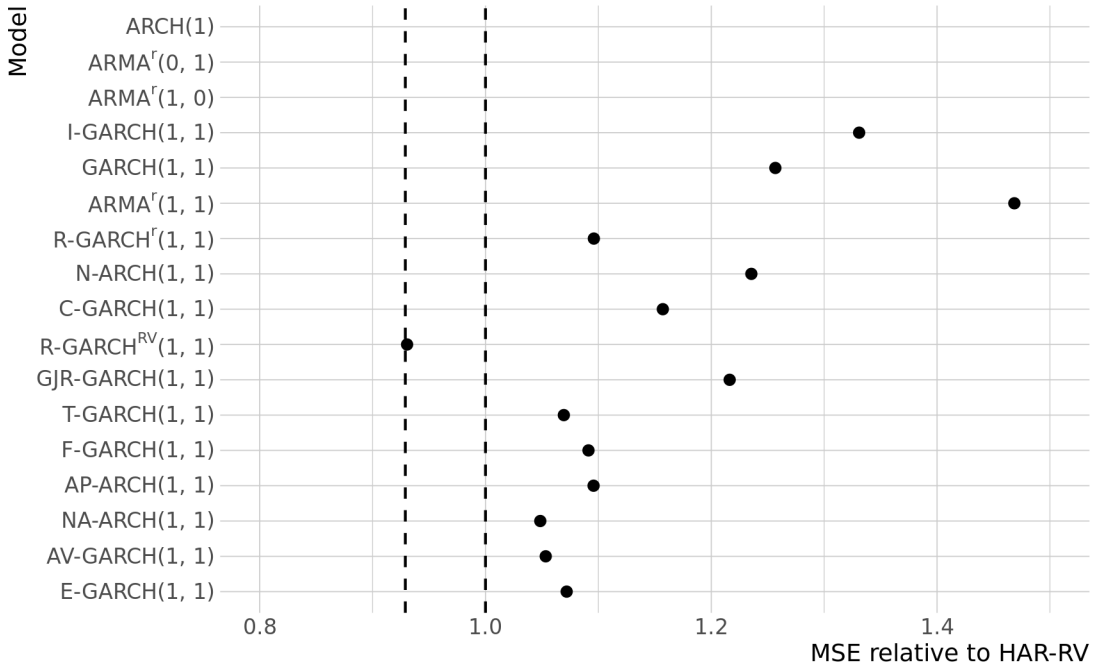
This figure compares the performance of all HAR models (section 4.2.3) and all ARMA models with realised variance (section 4.2.2) I tested, ordered by their performance. This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

similar, but the dispersion between results widens, which makes the `HAR-RVlog` the best individual model for the MSE during that period, and very close to the best model for the QLIKE.

Why does using RVlog bring such consistent improvements? Most likely, because using the logarithm generates more stable estimates of volatility. In the literature, this is often hinted at – RVlog is (unconditionally) roughly normally distributed (e.g., (Pong et al., 2004)) and Clements and Preve (2019) find that the `HAR-RVlog` yields similar improvements to other ways of getting more stable estimates, such as using the weighted least squares estimator. Furthermore, this is indirectly reflected in my results because when I do not replace negative HAR forecasts[34] (which are not shown in this thesis), there are three HAR models with an MSE 3 times higher than the MSE of the `HAR-RV`. For this reason, modelling the logarithm of realised volatility should probably be preferred.

[34] As I described in sections 4.2 and 4.3, I replace negative forecasts with the model's forecast of the previous day divided by two.

Figure 8: Out-of-sample MSE for GARCH and ARMA^r models (2005-2006)



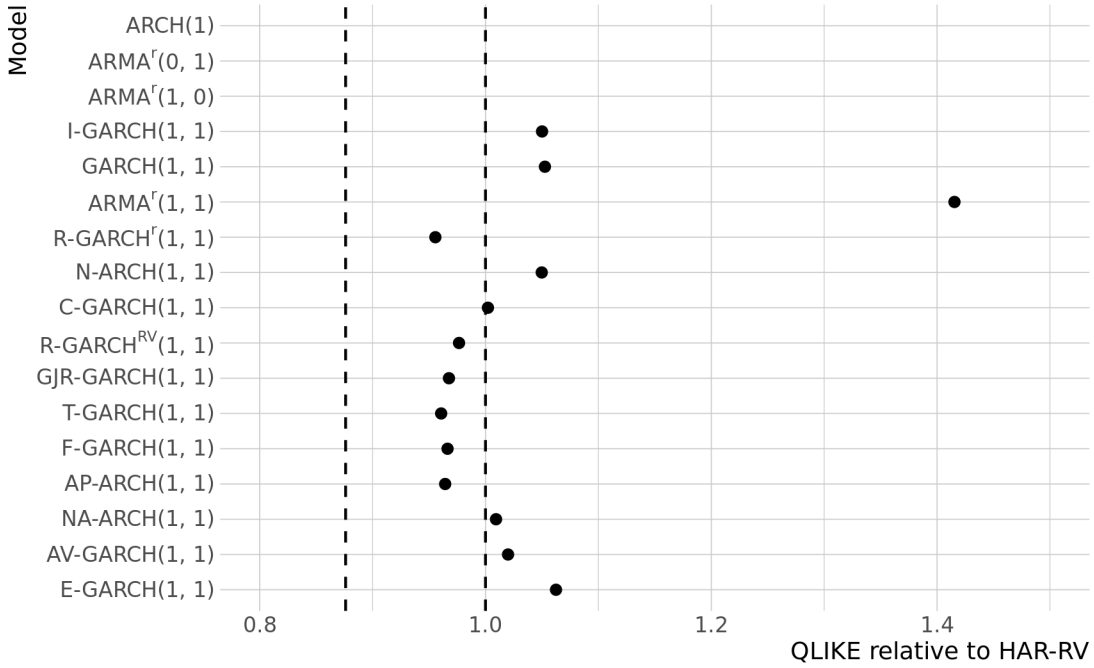
This figure compares the performance of all GARCH models (section 4.2.4) and all ARMA models with squared returns (section 4.2.2) I tested, ordered by MSE of the models in the period of 2007-2010. This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `HAR-RVlog` with a relative MSE of 0.93. Note that all these statistics are only calculated from observations in the period being considered in the figure (2005-2006). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

All other patterns in the data are extremely subtle: models with continuous (`C`) and jump (`J`) components are better, as are models with semi-variance (RV_t^+ and RV_t^-) which allow for the leverage effect. The inclusion of the quarticity component (`Q`) is even less consequential than other components. This does not allow me to say that it is expected to be less important in other data sets, as Bollerslev et al. (2016) shows evidence in favour of it with a larger sample size. That said, it is informative to know that this outperformance does not seem as consistent as the one reported by the `HAR-RVlog`, both in my results and the literature (Buccheri and Corsi, 2019; Clements and Preve, 2019).

5.2.3 GARCH and ARMA with squared returns

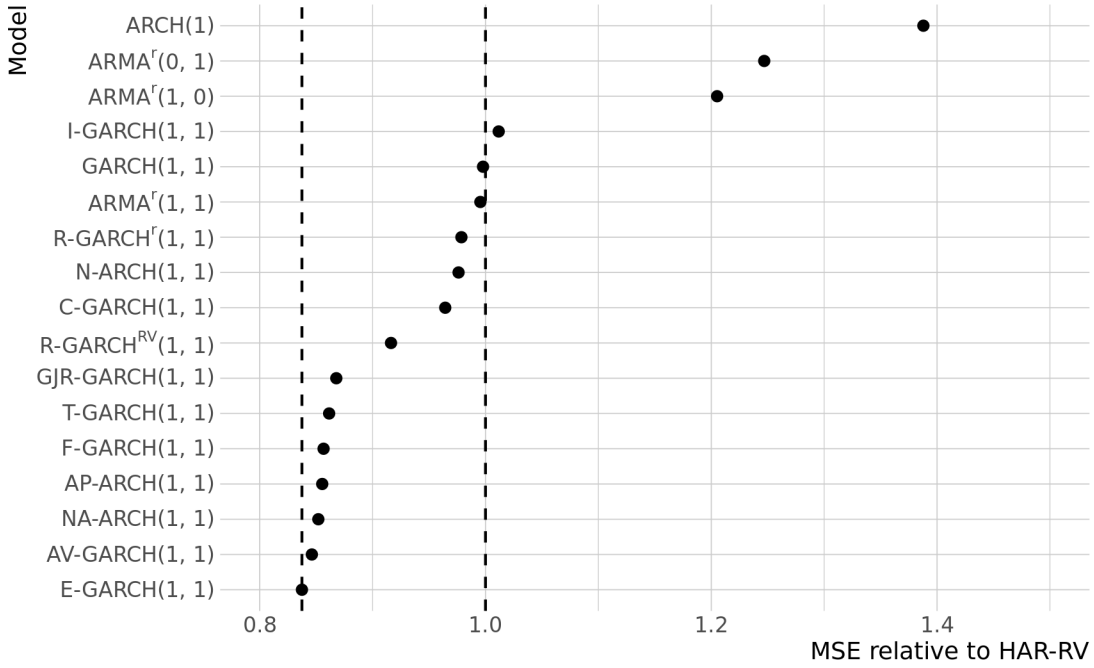
For individual models, GARCH models had the most surprising results: on average, they performed better than HAR models. This is, however, inconsistent because in the low volatility period of 2005-2006 it had a relatively low MSE, but a relatively high QLIKE (Fig-

Figure 9: Out-of-sample QLIKE for GARCH and ARMA^{r} models (2005-2006)



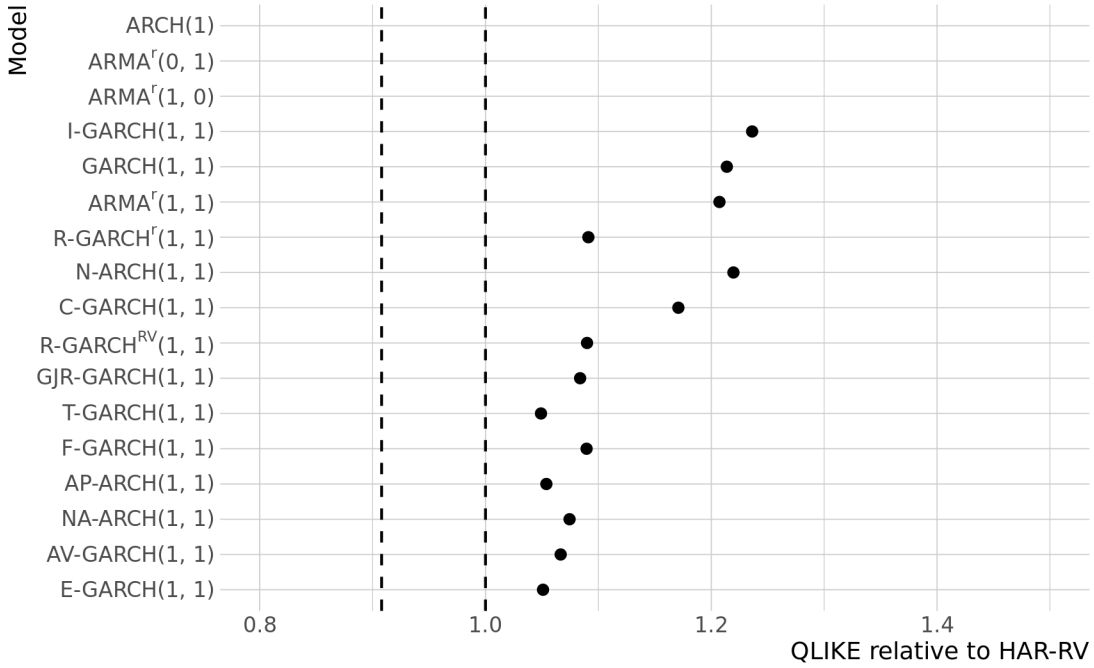
This figure compares the performance of all GARCH models (section 4.2.4) and all ARMA models with squared returns (section 4.2.2) I tested, ordered by MSE of the models in the period of 2007-2010. This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `QS-HAR-RV` with a relative QLIKE of 0.88. Note that all these statistics are only calculated from observations in the period being considered in the figure (2005-2006). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

Figure 10: Out-of-sample MSE for GARCH and ARMA^{r} models (2007-2010)



This figure compares the performance of all GARCH models (section 4.2.4) and all ARMA models with squared returns (section 4.2.2) I tested, ordered by the MSE of the models in the period of 2007-2010. This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH`$(1,1)$ with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix □.

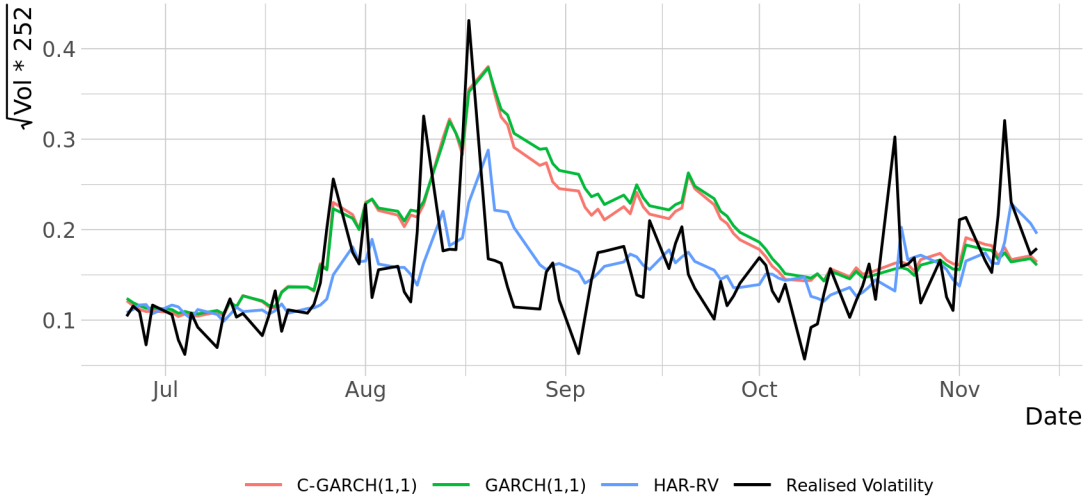
Figure 11: Out-of-sample QLIKE for GARCH and ARMA^{r} models (2007-2010)



This figure compares the performance of all GARCH models (section 4.2.4) and all ARMA models with squared returns (section 4.2.2) I tested, ordered by the MSE of the models in the period of 2007-2010. This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

ures 8 and 9); while in the high volatility period of 2007-2010 it was the opposite – a relatively high MSE, but low QLIKE (Figures 10 and 11). Through visual inspection (Figure 12 shows one prototypical example), it is possible to observe a pattern: GARCH models tend to have too much persistence after a large shock. This was previously observed with the 1987 crash Schwert (1989), and the `C-GARCH` was constructed to correct such problems, but in my sample this does not happen successfully. The GARCH variants that perform better are the ones that allow for the leverage effect, which is corroborated by Hansen and Lunde (2005) and Awartani and Corradi (2005). The inclusion of the realised volatility (through the `R-GARCH`^{RV}) was not consequential in almost every case.

Figure 12: Example: GARCH vs HAR after a volatility shock

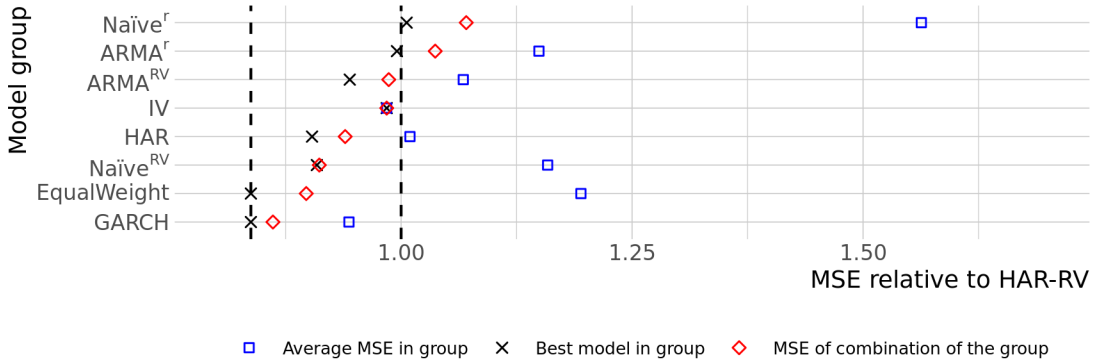


The realised volatility (from 2006-07-25 to 2007-11-13) is shown in black, and the one-day-ahead volatility forecasts by other models ("Vol") with 3 different colors. This is the typical pattern with GARCH models in my sample – they tend to not mean-revert as quickly as they should. The `C-GARCH(1,1)` was made to solve this issue, so it is shown in the comparison.

Despite my results, I do not expect that GARCH models do as well as HAR models in the long run. There is enough evidence in the literature to suggest that GARCH models sometimes underperform more heavily relative to models with realised volatility (e.g., Corsi (2004)), and my sample is mostly composed of a period of very high volatility by historical standards. But if GARCH models do perform satisfactorily during such high volatility periods, this is still a very meaningful result, as these are the periods with a disproportionately high impact on risk management. This conclusion could also be applied to exponential smoothing models with squared returns that, despite being slightly inferior in my sample, seem to have equivalent performance to GARCH models in the literature[35] (Andersen et al,

[35]In most of the literature, the exponential smoothing model with $\texttt{beta} = 0.94$ is referred to as the "RiskMetrics" model (Morgan et al, 1996).

Figure 13: Out-of-sample MSE for Grouping and Equal Weight combinations (2007-2010)



Within each category, I mark the performance of the best individual model with a cross (e.g., for GARCH models, the cross represents the performance of the `E-GARCH`$(1,1)$ in this case). The blue squares represent the average performance of the individual models. And the red diamonds are the "Grouping" (section 4.3.1) models. They are obtained by averaging all forecasts of their respective model categories – the forecasts for the "GARCH" model group, for example, are obtained by averaging the forecasts of the 14 individual GARCH models. I also include the `EqualWeight` as a point of reference, which averages all (74) individual model forecasts.
This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH`$(1,1)$ with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

2003; Corsi, 2004).

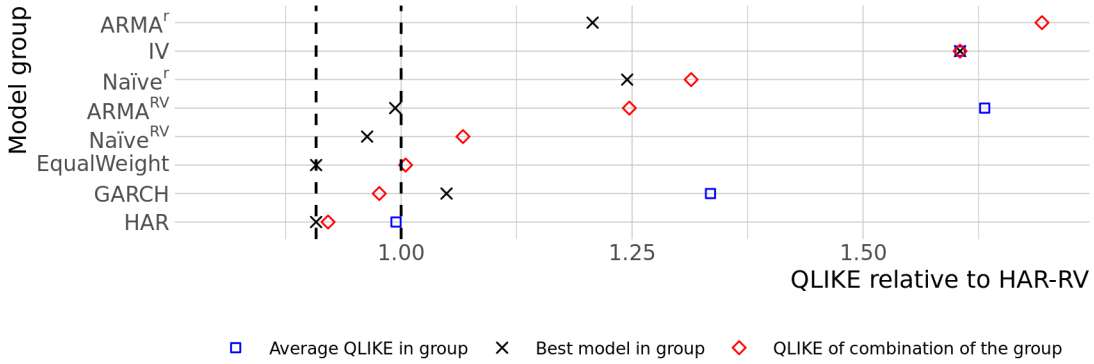
5.3 Model combinations

5.3.1 Grouping and equal weighting models

I first look at averaging the forecasts within each category (Figures 13 and 14). There, we can observe a stylized fact from the model combinations literature (Timmermann, 2006): the performance of the *average forecast* (red diamonds) is consistently better than the performance of the *average model* (blue squares), although it rarely surpasses the best model within that category (black crosses). The improvements tend to get larger with worse model categories, which is why the improvements for the QLIKE are larger. Worse models benefit more from being combined because the reasons for the improvements (such as models not taking advantage of all information, or models having noisy estimates) are more often present in naïve models, and models that use squared returns.

I also combine multiple model categories (Figures 15 and 16), similarly to Makridakis and Winkler (1983) – for instance, instead of just pooling HAR forecasts, I can pool HAR and GARCH forecasts. As I combine more models from different categories, the average and the variance of the losses decrease, although the effect is smaller with the QLIKE – there, as I

Figure 14: Out-of-sample QLIKE for Grouping and Equal Weight combinations (2007-2010)



Within each category, I mark the performance of the best individual model with a cross (e.g., for HAR models, the cross represents the performance of the `HAR-C-J` in this case). The blue squares represent the average performance of the individual models. And the red diamonds are the "Grouping" (section 4.3.1) models. They are obtained by averaging all forecasts of their respective model categories – the forecasts for the "GARCH" model group, for example, are obtained by averaging the forecasts of the 14 individual GARCH models. I also include the `EqualWeight` as a point of reference, which averages all (74) individual model forecasts.
This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

add more model categories, the probability of using a combination that improves on the `HAR-RV` gets smaller.

To understand why this is, it helps to look at the results in Figures 17 and 18: combinations of HAR and/or GARCH, which are the best models, are sufficient to bring the reduction in variance, while also not deteriorating the expected performance. Hence, inferior models do not contribute to improved performance, although the differences are very small (as can be seen by the equal weight scheme).

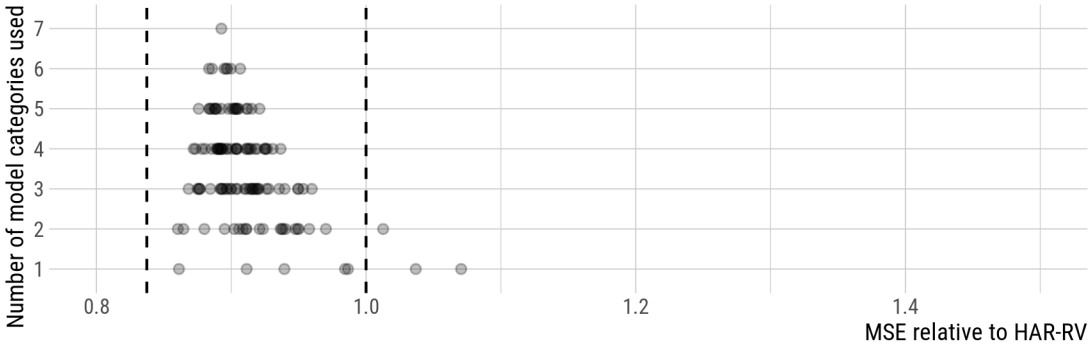
5.3.2 Trimming

From the previous sub-section, it becomes clear that combining the best categories leads to better results. But instead of trying to combine the best models by choosing the right categories, we can explicitly try to achieve that goal by trimming the worst forecasts (Figures 19 and 20).

Once again, the improvements are very small compared to the equal weighting scheme, and it sometimes does worse than combinations of GARCH and/or HAR. But the results are a bit more consistent for both loss functions, especially for $1.1 \leqslant k \leqslant 1.4$, which excludes an average of roughly 10 to 40 individual models per day (out of the 74).

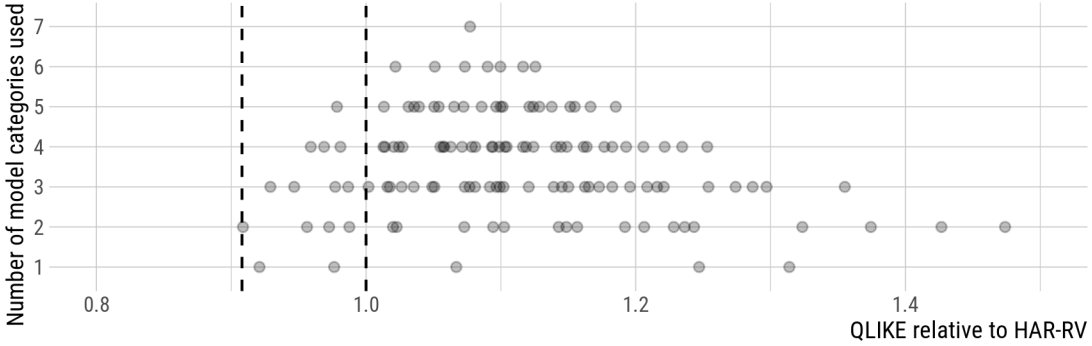
So, although trimming has the best average performance across every specification tested

Figure 15: Out-of-sample MSE for Grouping combinations (2007-2010)



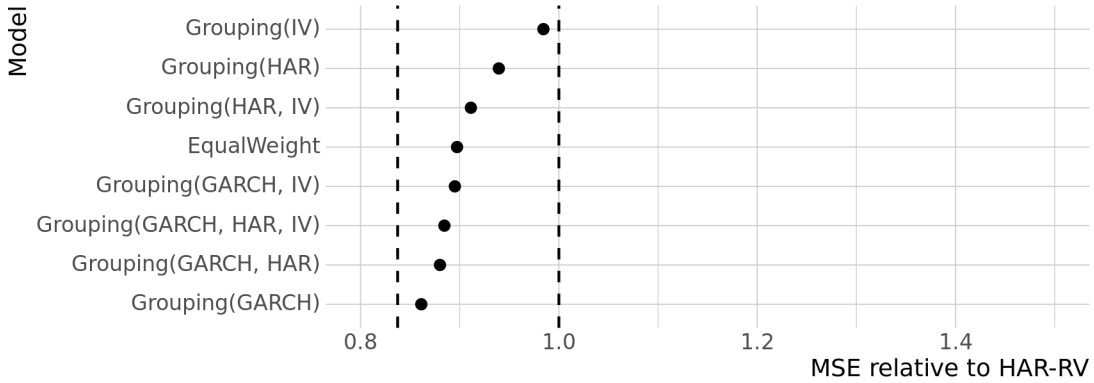
Each black point is a "Grouping" model (section 4.3.1), which is a way to combine models. The row with 1 model category shows the same data as the red diamonds in Figure 13. The row with 2 model categories shows all the possible Grouping combinations that can be made with the 7 individual model categories (HAR with GARCH, HAR with IV, etc.). The other rows have the same interpretation – they give the performance of Grouping models that can be made with n individual model categories.
This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH`$(1,1)$ with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

Figure 16: Out-of-sample QLIKE for Grouping combinations (2007-2010)



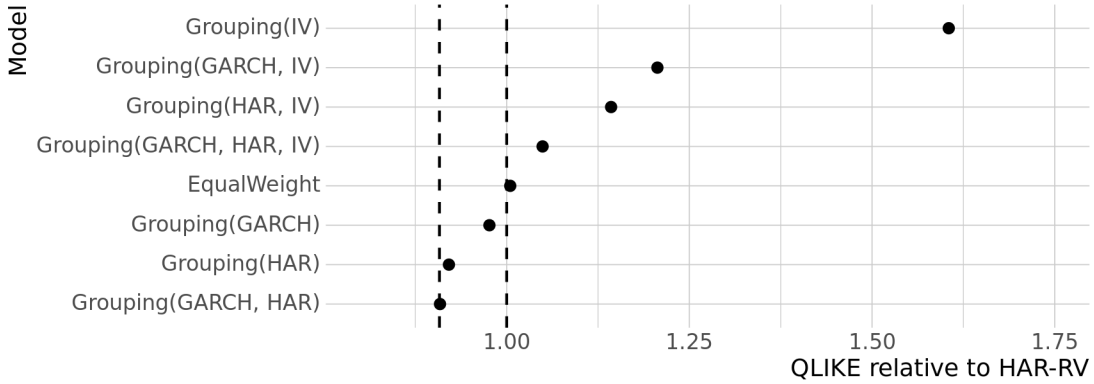
Each black point is a "Grouping" (section 4.3.1) model, which is a way to combine models. The row with 1 model category shows the same data as the red diamonds in Figure 14. The row with 2 model categories shows all the possible Grouping combinations that can be made with the 7 individual model categories (HAR with GARCH, HAR with IV, etc.). The other rows have the same interpretation – they give the performance of Grouping models that can be made with n individual model categories.
This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

Figure 17: Out-of-sample MSE for selected Grouping models (2007-2010)



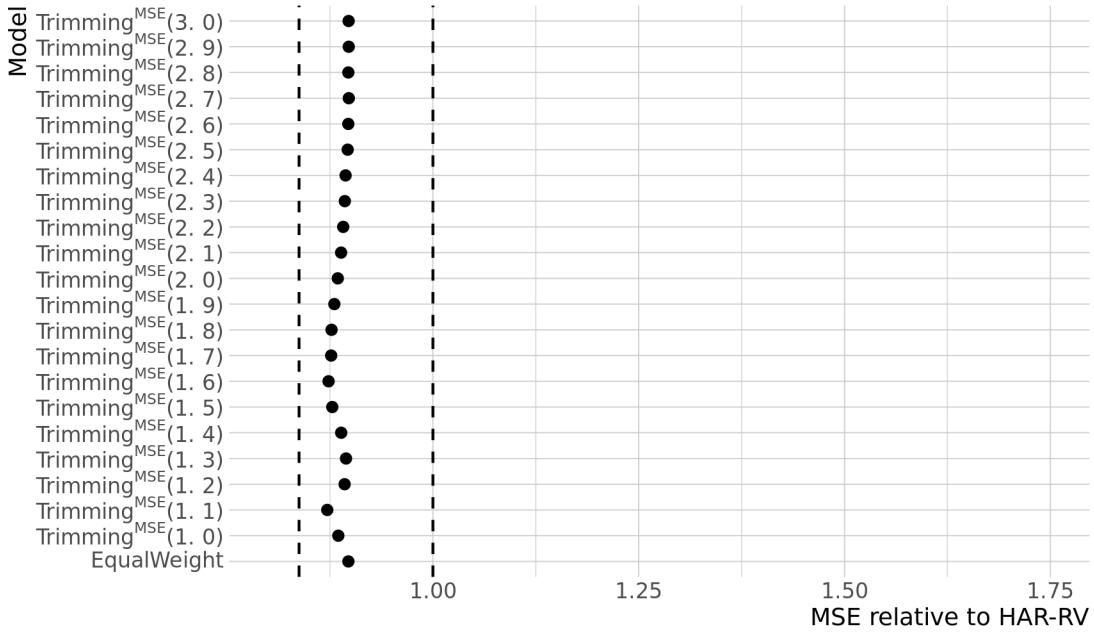
Out of all the Grouping models presented in Figure 15, I present the ones containing all possible combinations of GARCH, HAR, and IV models – a priori, I considered those to be the models that would most benefit from being combined.
This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH`$(1,1)$ with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix O.

Figure 18: Out-of-sample QLIKE for selected Grouping models (2007-2010)



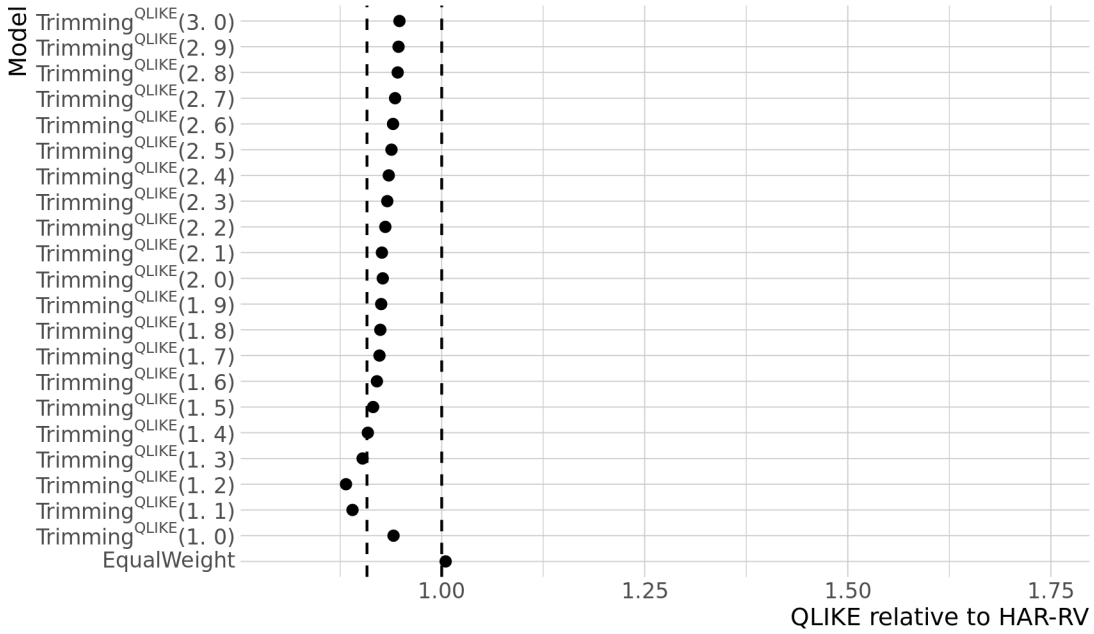
Out of all the Grouping models presented in Figure 16, I present the ones containing all possible combinations of GARCH, HAR, and IV models – a priori, I considered those to be the models that would most benefit from being combined.
This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix O.

Figure 19: Out-of-sample MSE for trimming and Equal Weight combinations (2007-2010)



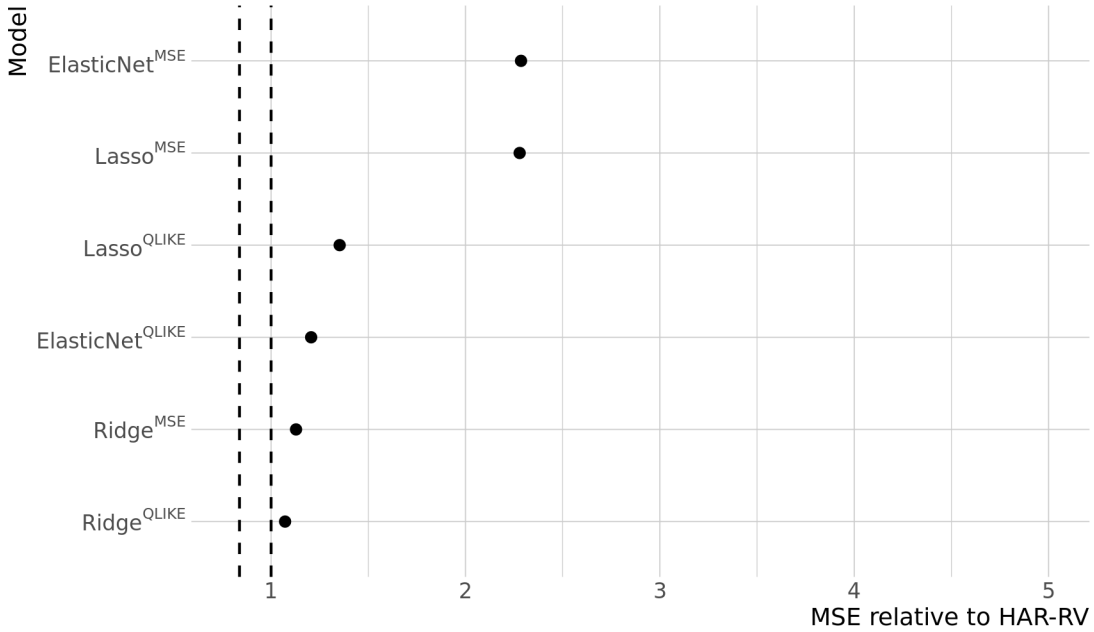
Trimming$^{MSE}(k)$ excludes all the models whose past MSE is above the k threshold. That threshold, given by k, is the ratio between that past MSE, and the past MSE of the model with the lowest MSE. If $k = 2$, for example, and the best model had an MSE of $1.5e-07$, the trimming procedure excludes any models with MSE $> 3e-07$.
This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH`$(1,1)$ with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix ◻.

Figure 20: Out-of-sample QLIKE for trimming and Equal Weight combinations (2007-2010)



$\text{Trimming}^{\text{QLIKE}}(k)$ excludes all the models whose past MSE is above the k threshold. That threshold, given by k, is the ratio between that past MSE, and the past QLIKE of the model with the lowest QLIKE. If $k = 2$, for example, and the best model had a QLIKE of $1.5e-07$, the trimming procedure excludes any models with QLIKE $> 3e-07$.
This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix □.

Figure 21: Out-of-sample MSE for regularization regressions (2007-2010)



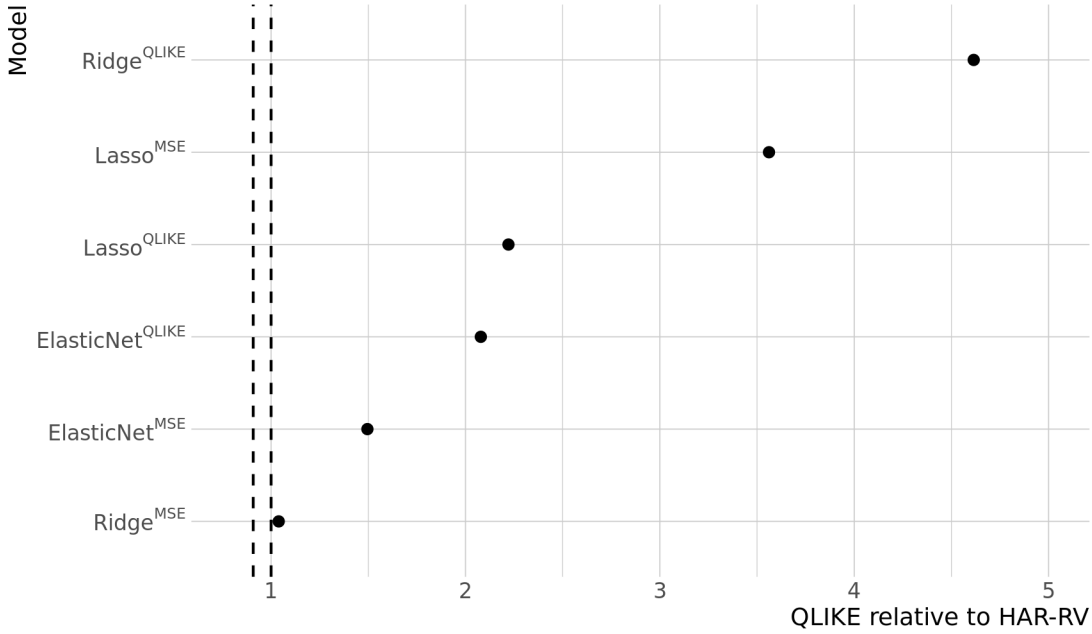
This figure compares the performance of all Regularization models (section 4.3.3) I tested, ordered by the MSE of the models in the period of 2007-2010.
This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH`$(1,1)$ with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

(Figures 2 and 3) the main conclusion is that any (simple) form of combining forecasts leads to more consistent results, with a similar MSE and QLIKE to the (ex-post) optimal model. This is corroborated by Timmermann (2006), which also finds that trimming is the better option for making forecast combinations.

5.3.3 Regularization: Ridge, LASSO and Elastic Net

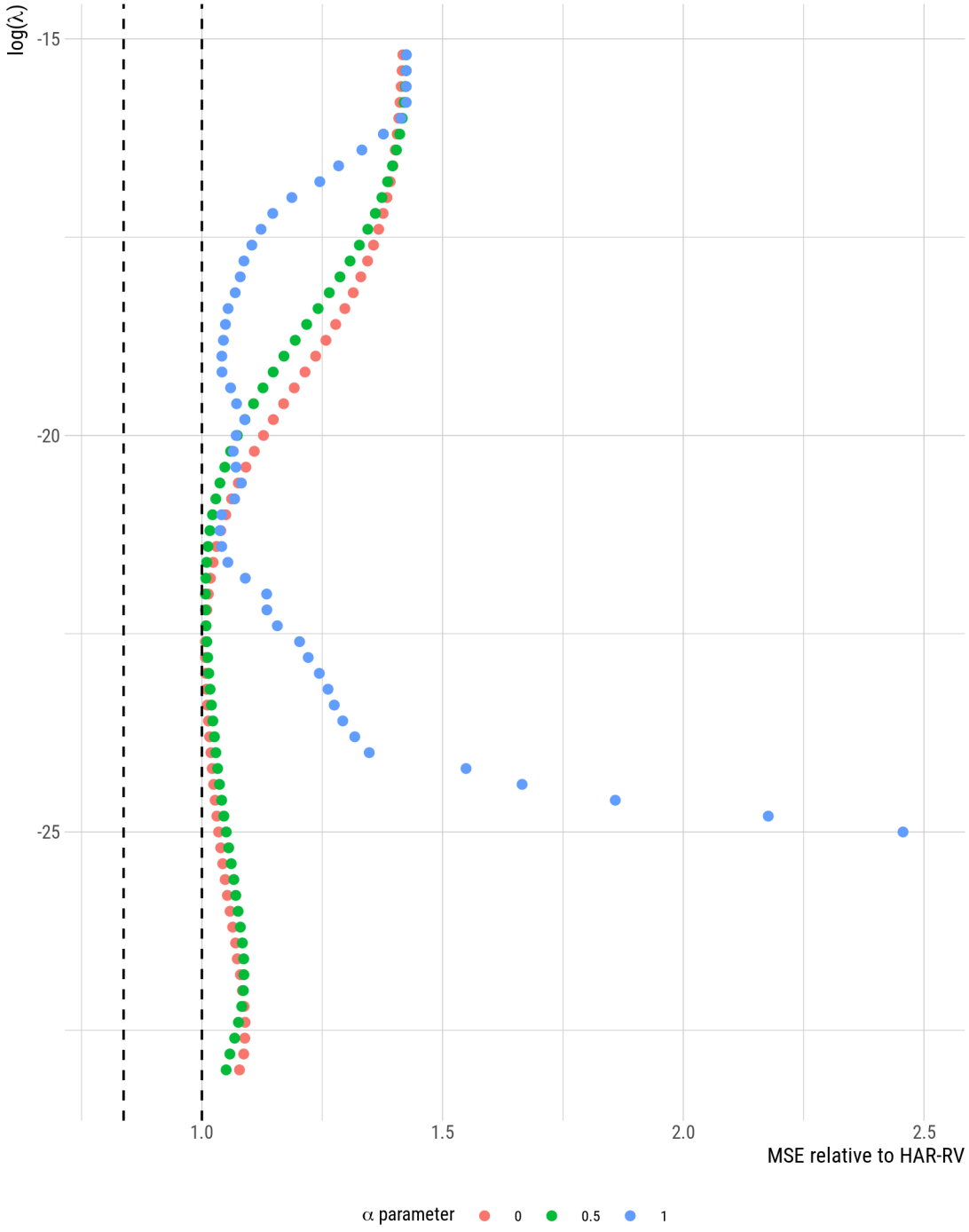
I can additionally improve forecasts by exploiting complex relationships that exist between individual models. This complexity introduces the additional difficulty of correctly estimating hyperparameters due to overfitting, which is difficult to deal with in the estimation process – as I mentioned in section 4.3.3, applying the one-standard-deviation rule made the procedure select intercept-only models. Notice in Figures 21 and 22 that the effect of overfitting is noticeable: the models optimized on the MSE do better in terms of the QLIKE, and the models optimized on the QLIKE do better in terms of the MSE. Lastly, note that the ridge does predictably better than all other types of regularization techniques for the

Figure 22: Out-of-sample QLIKE for regularization regressions (2007-2010)



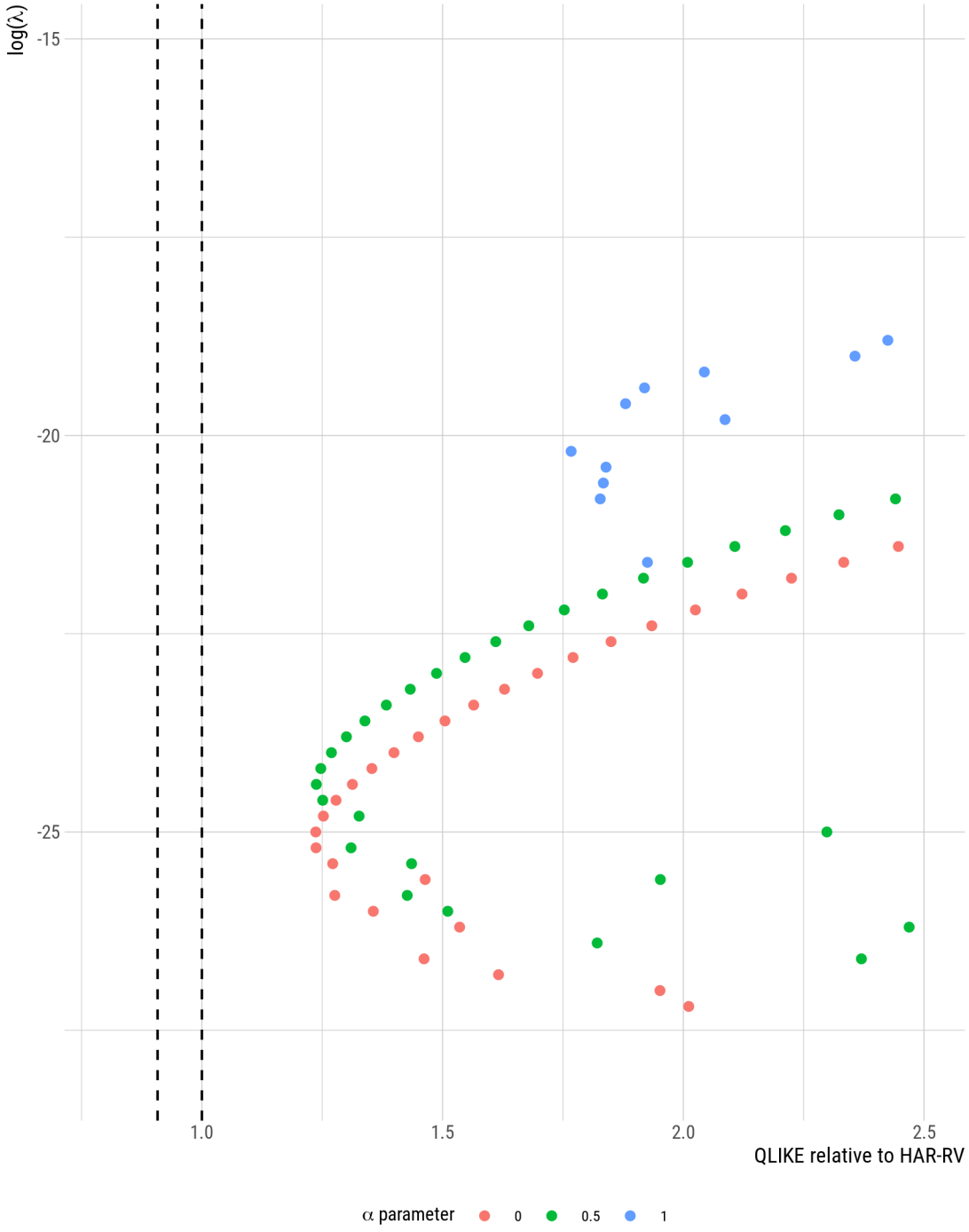
This figure compares the performance of all Regularization models (section 4.3.3) I tested, ordered by the MSE of the models in the period of 2007-2010.
This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

Figure 23: Out-of-sample MSE for regularization regressions with fixed λ (2007-2010)



Whereas Figures 21 and 22 show Regularization models that dynamically choose the shrinkage parameter (λ), this figure shows the performance of those models when λ is fixed. The three α values shown, 0, 0.5 and 1, correspond to the Ridge regression, Elastic Net, and Lasso methods.
This is a relative comparison: every mean squared error is divided by the mean squared error of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative MSE – in this case, the `E-GARCH`$(1,1)$ with a relative MSE of 0.84. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix D.

Figure 24: Out-of-sample QLIKE for regularization regressions with fixed λ (2007-2010)



Whereas Figures 21 and 22 show Regularization models that dynamically choose the shrinkage parameter (λ), this figure shows the performance of those models when λ is fixed. The three α values shown, 0, 0.5 and 1, correspond to the Ridge regression, Elastic Net, and Lasso methods.
This is a relative comparison: every quasi-likelihood is divided by the quasi-likelihood of the `HAR-RV`, so a performance of 1.0 (marked with the rightmost dotted vertical line) means that the performance is equal to the one from the HAR-RV. The leftmost dotted vertical line marks the individual model with the best relative QLIKE – in this case, the `HAR-C-J` with a relative QLIKE of 0.91. Note that all these statistics are only calculated from observations in the period being considered in the figure (2007-2010). Some results may be missing from the figure because they could not fit in the figure. All the results, along with the corresponding MCS p-values, are expressed on the tables from Appendix C.

reasons I mentioned in section 4.3.3, but the results are still inferior and more variable than the ones obtained in simpler weighting schemes, so regularization adds no value (at least in the way I applied it).

To have a better sense of how these regularization methods were overfitting, I display results for $\alpha = 0$ (ridge), $\texttt{alpha} = 0.5$ (elastic net), and $\alpha = 1$ (lasso) with all the possible shrinkage parameters and their respective (out-of-sample) losses in Figures 23 and 24. For the MSE, the best shrinkage parameter only matches the `HAR-RV`; and for the QLIKE, the results are worse and more unstable. The figures also show that increasing the α hyperparameter (which puts a larger focus on excluding predictors) leads to identical performance so long as you increase the shrinkage parameter accordingly and, roughly speaking, set $\alpha < 0.9$ (the results for all α values used are not shown in this thesis).

This ex-post analysis does not necessarily represent best-case-scenario in terms of performance – if the shrinkage parameter decreases in magnitude as the standard errors get smaller, the loss should also be smaller than when using a fixed shrinkage parameter (this happens, for example, with the ridge regression for the QLIKE). Such an improvement seems, however, hard to obtain and comes at the cost of considerable variance in the results.

Timmermann (2006) reviews the literature and finds that shrinkage methods are often more successful than in my case; but his review also shows that the results are "quite sensitive to the shrinkage parameters" (page 184) and that trimming often does better, so there is, in large part, concordance between what I find and the literature.

6 Discussion

Broadly speaking, there are two ways to improve out-of-sample forecasts: we can use more data in our model, or we can get better estimates with the data we already have. The new data has come in the form of utilizing multivariate models (Liu, 2009), new realised measures with intraday returns and tick data (Liu et al., 2015), and macroeconomic or alternative data sets (Bollerslev et al., 2009; Oliveira et al., 2017; Wang et al., 2006). For short-term forecasts (roughly 1- to 10-day-forecasts), only the advent of realised measures has substantially impacted performance, and new measures tend to offer very slight improvements, as I demonstrate in my empirical results and the Literature Review (section 2.2).

To get better estimates with the data available, there are numerous methods in the literature such as the use of variations on GARCH models (Hansen and Lunde, 2005), different estimators and link functions (Clements and Preve, 2019), outlier removal (Franses and Ghijsels, 1999), model comparisons (Timmermann, 2006), among others. Barring some exceptions, few of these methods offer consistent improvements in forecast performance. The lack of improvements is even more salient for longer-term forecasts (close to, or larger than, 20-day-ahead forecasts) since they mostly dominated by implied volatility and models with realised volatility also seem adequate at making long-term predictions (Christensen and Prabhala, 1998; Pong et al., 2004).

So if all alternatives to improve forecasts have had limited success, it is important to prove that small increments are correctly measured and useful in applied settings. But almost no papers do this. One of the few articles is by Neely (2009) that tests how the delta hedging practice is improved by augmenting implied volatility with statistical models; he finds that the tracking error is not economically significant and rarely statistically significant, and that augmenting IV with models using squared returns was not consistently better than augmenting IV with models using realised variance. In the context of portfolio optimization, Liu (2009) find that intraday data improves on squared returns, but only when rebalancing the portfolio every day.

So if the benefits given by these models are limited and often lack of economic significance, it puts into question the usefulness of the paradigm used in the literature that I analyse, which tends to measure progress based on the loss of point forecasts made by univariate models. At a minimum, researchers must be more conservative when equating improvements in a loss function to improvements in applied settings. And to advance the knowledge in this area, researchers must test economic significance more often – in settings such as delta hedging or some other kind of risk management (e.g., (Liu, 2009)); determining risk premia (Bollerslev et al., 2009); methodological improvements (Bollerslev et al. (1992) give a few examples of applications related to this); focus on improving forecasts just with

squared returns, since most practitioners do not have access to intraday data (Clements and Preve (2019), for example, show that a range-based volatility proxy with daily high-low-close data performs well); among others.

To assess statistical significance, I find the MCS test to generally be insufficient when there are many models relative to the number of observations. Consider the anecdote that when I use the MCS test with the QLIKE with just the 74 individual models in my sample, 4 models are excluded from the set of superior models; but when I also include model combinations (another 176 models), only 3 models are excluded in total, even though almost all the model combinations have close-to-optimal performance. This behaviour is caused by a large loss of power, which is itself mainly caused by the test trying to control the rate of at least one falsely-rejected model (familywise error rate, or FWER).

The alternatives covered in the Literature Review (section 2.5) have similar issues, in that they often try to control for the FWER, and they can lack sufficient flexibility for the particular setting (e.g., not allowing for nested models). A worthwhile alternative is to use Bayesian multilevel models (Gelman, Hill, and Yajima, 2012) that has the advantages of being more interpretable, flexible and easily extensible, and allows you to incorporate knowledge from past research, which is valuable when the literature is as extensive as the one about volatility forecasting.

7 Conclusion

I compare the main univariate models in the financial market volatility literature, and test model combinations. With individual models, the GARCH, HAR, `ARMA(1,1)`, and exponential smoothing models work similarly well, though the models using realised variance are more consistent. GARCH and HAR models are better when they allow for the leverage effect; HAR models are additionally improved by the simultaneous inclusion of continuous and jump realised measures, and by the use of the logarithm of realised variance (RVlog); and exponential smoothing models perform well when their parameter allows for high or intermediate volatility persistence (i.e., $0.5 \leqslant \beta \leqslant 0.95$).

Model combinations have, on average, a lower loss than individual models: for any given model category, the loss of individual models is *more variable* than the model combination made with the models from that category; the loss of individual models is also *higher* than the model combination; but the loss of the *best individual model* of a category is lower than the model combination made with models from that category. This effect is stronger for the mean squared error (MSE) than for the quasi-likelihood (QLIKE); for simple model combination weighting schemes (i.e., everything I tested except regularization); and for model combinations that exclude bad performers directly (by trimming) or indirectly (by only combining the best-performing models to begin with, such as only combining HAR and GARCH models).

All of these differences in performance are very small. To assess their significance, I use the model confidence set (MCS) test (Hansen et al, 2011), which I consider to be the most adequate test in the literature. At the same time, it pursues the wrong goal, which is to control the rate at which at least one superior model is excluded from the set of superior models (SSM). This issue, among other things, reduces the power of the test substantially, so it does not reject many models that are clearly inferior – with α set to 0.2, for the QLIKE, only 3 out of the 250 models tested are not included in the SSM; for the MSE, no model is excluded because models are closer to each other on that loss function. So, despite the lack of significant results, I find it probable that many of the observed patterns mentioned above will be seen in other samples, especially for those that have also been documented in the literature[36] (e.g., Hansen and Lunde (2005) and Awartani and Corradi (2005) also find that GARCH models are better when they allow for the leverage effect, and I essentially replicate most of the main findings of Timmermann (2006) about model combinations).

In light of the research that applies these models in applied economic settings, it is almost surely the case that, in practice, using any reasonably adequate model would give

[36] One limitation of this study is that the index used (FTSE100) is probably correlated with other widely studied indexes so, to some extent, I would expect to see similar results.

identical benefits under most circumstances. Since many avenues to improve performance have already been tried with little success, future research needs to measure progress with metrics directly related to economic applications that models are supposed to improve, as loss functions do not translate well into those.

References

Aiolfi, M. and C. A. Favero (2005). Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting 24*(4), 233–254.

Andersen, T. G. and T. Bollerslev (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, 885–905.

Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold (2005). Volatility forecasting. Technical report, National Bureau of Economic Research.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The review of economics and statistics 89*(4), 701–720.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The distribution of realized stock return volatility. *Journal of financial economics 61*(1), 43–76.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (1999). Realized volatility and correlation. *LN Stern School of Finance Department Working Paper 24*.

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

Arlot, S., A. Celisse, et al. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys 4*, 40–79.

Audrino, F. and S. D. Knaus (2016). Lassoing the har model: A model selection perspective on realized volatility dynamics. *Econometric Reviews 35*(8-10), 1485–1521.

Awartani, B. M. and V. Corradi (2005). Predicting the volatility of the s&p-500 stock index via garch models: the role of asymmetries. *International Journal of forecasting 21*(1), 167–183.

Baillie, R. T., T. Bollerslev, and H. O. Mikkelsen (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of econometrics 74*(1), 3–30.

Bandi, F. M. and J. R. Russell (2008). Microstructure noise, realized variance, and optimal sampling. *The Review of Economic Studies 75*(2), 339–369.

Barndorff-Nielsen, O. E., S. Kinnebrock, and N. Shephard (2008). Measuring downside risk-realised semivariance. *CREATES Research Paper* (2008-42).

Barndorff-Nielsen, O. E. and N. Shephard (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(2), 253–280.

Barndorff-Nielsen, O. E. and N. Shephard (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of financial econometrics 2*(1), 1–37.

Becker, R. and A. E. Clements (2008). Are combination forecasts of s&p 500 volatility statistically superior? *International Journal of Forecasting 24*(1), 122–133.

Becker, R., A. E. Clements, and S. I. White (2007). Does implied volatility provide any information beyond that captured in model-based volatility forecasts? *Journal of Banking & Finance 31*(8), 2535–2549.

Bergmeir, C. and J. M. Benítez (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences 191*, 192–213.

Bergmeir, C., R. J. Hyndman, and B. Koo (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis 120*, 70–83.

Bernardi, M. and L. Catania (2018). The model confidence set package for r. *International Journal of Computational Economics and Econometrics 8*(2), 144–158.

Black, F. (1976). Studies of stock price volatility changes. In *Proceedings of the 1976 Meeting of the Business and Economic Statistics Section, American Statistical Association*, Washington DC, pp. 177–181.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics 31*(3), 307–327.

Bollerslev, T., R. Y. Chou, and K. F. Kroner (1992). Arch modeling in finance: A review of the theory and empirical evidence. *Journal of econometrics 52*(1-2), 5–59.

Bollerslev, T., J. Litvinova, and G. Tauchen (2006). Leverage and volatility feedback effects in high-frequency data. *Journal of Financial Econometrics 4*(3), 353–384.

Bollerslev, T., A. J. Patton, and R. Quaedvlieg (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics 192*(1), 1–18.

Bollerslev, T., G. Tauchen, and H. Zhou (2009). Expected stock returns and variance risk premia. *The Review of Financial Studies 22*(11), 4463–4492.

Box, G. E. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological) 26*(2), 211–243.

Brailsford, T. J. and R. W. Faff (1996). An evaluation of volatility forecasting techniques. *Journal of Banking & Finance 20*(3), 419–438.

Brooks, C. and G. Persand (2003). Volatility forecasting for risk management. *Journal of forecasting 22*(1), 1–22.

Buccheri, G. and F. Corsi (2019). Hark the shark: Realized volatility modelling with measurement errors and nonlinear dependencies. *Available at SSRN 3089929*.

Cawley, G. C. and N. L. Talbot (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research 11*(Jul), 2079–2107.

Chernov, M. (2007). On the role of risk premia in volatility forecasting. *Journal of Business & Economic Statistics 25*(4), 411–426.

Christensen, B. J. and N. R. Prabhala (1998). The relation between implied and realized volatility. *Journal of financial economics 50*(2), 125–150.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting 5*(4), 559–583.

Clements, A. and D. Preve (2019). A practical guide to harnessing the har volatility model. *Available at SSRN 3369484*.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues.

Corsi, F. (2004). A simple long memory model of realized volatility. *Available at SSRN 626064*.

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

Corsi, F., D. Pirino, and R. Reno (2010). Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics 159*(2), 276–288.

Corsi, F. and R. Renò (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics 30*(3), 368–380.

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

Ding, Z., C. W. Granger, and R. F. Engle (1993). A long memory property of stock market returns and a new model. *Journal of empirical finance 1*(1), 83–106.

Donaldson, R. G. and M. J. Kamstra (2005). Volatility forecasts, trading volume, and the arch versus option-implied volatility trade-off. *Journal of Financial Research 28*(4), 519–538.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association 56*(293), 52–64.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.

Engle, R. F. and T. Bollerslev (1986). Modelling the persistence of conditional variances. *Econometric reviews 5*(1), 1–50.

Engle, R. F. and G. Lee (1999). A long-run and short-run component model of stock return volatility. *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive WJ Granger*, 475–497.

Engle, R. F. and V. K. Ng (1993). Measuring and testing the impact of news on volatility. *The journal of finance 48*(5), 1749–1778.

Figlewski, S. (1997). Forecasting volatility. *Financial markets, institutions & instruments 6*(1), 1–88.

Franses, P. H. and H. Ghijsels (1999). Additive outliers, garch and forecasting volatility. *International Journal of Forecasting 15*(1), 1–9.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

FTSE Russell (2020, March). Ftse uk index series. Technical Report 14.1, FTSE International, Ltd.

Gelman, A., J. Hill, and M. Yajima (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness 5*(2), 189–211.

Gelman, A. and H. Stern (2006). The difference between significant and not significant is not itself statistically significant. *The American Statistician 60*(4), 328–331.

Ghalanos, A. (2014). *rugarch: Univariate GARCH models.* R package version 1.4-0.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance 48*(5), 1779–1801.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics 23*(4), 365–380.

Hansen, P. R., Z. Huang, and H. H. Shek (2012). Realized garch: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics 27*(6), 877–906.

Hansen, P. R. and A. Lunde (2005). A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of applied econometrics 20*(7), 873–889.

Hansen, P. R. and A. Lunde (2006). Consistent ranking of volatility models. *Journal of Econometrics 131*(1-2), 97–121.

Hansen, P. R., A. Lunde, and J. M. Nason (2003). Choosing the best volatility models: the model confidence set approach. *Oxford Bulletin of Economics and Statistics 65*, 839–861.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hentschel, L. (1995). All in the family nesting symmetric and asymmetric garch models. *Journal of Financial Economics 39*(1), 71–104.

Higgins, M. L. and A. K. Bera (1992). A class of nonlinear arch models. *International Economic Review*, 137–158.

Hill, C. and B. McCullough (2019). On the accuracy of garch estimation in r packages. *Econometric Research in Finance 4*(2), 133–156.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Hou, K., C. Xue, and L. Zhang (2017). Replicating anomalies. *The Review of Financial Studies*.

Hyndman, R. J., G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, and E. Wang (2020). Package forecast. *Online] https://cran. r-project. org/web/packages/forecast/forecast. pdf* .

Kambouroudis, D. S., D. G. McMillan, and K. Tsakou (2016). Forecasting stock return volatility: a comparison of garch, implied volatility, and realized volatility models. *Journal of Futures Markets 36*(12), 1127–1163.

Lamoureux, C. G. and W. D. Lastrapes (1993). Forecasting stock-return variance: Toward an understanding of stochastic implied volatilities. *The Review of Financial Studies 6*(2), 293–326.

LeBaron, B. et al. (2001). Stochastic volatility as a simple generator of apparent financial power laws and long memory. *Quantitative Finance 1*(6), 621–631.

Liu, L. Y., A. J. Patton, and K. Sheppard (2015). Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. *Journal of Econometrics 187*(1), 293–311.

Liu, Q. (2009). On portfolio optimization: How and when do we benefit from high-frequency data? *Journal of Applied Econometrics 24*(4), 560–582.

Lo, A. W. and A. C. MacKinlay (1990). Data-snooping biases in tests of financial asset pricing models. *The Review of Financial Studies 3*(3), 431–467.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting 36*(1), 54–74.

Makridakis, S. and R. L. Winkler (1983). Averages of forecasts: Some empirical results. *Management science 29*(9), 987–996.

Mancini, C. (2009). Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics 36*(2), 270–296.

Martens, M., D. Van Dijk, and M. De Pooter (2009). Forecasting s&p 500 volatility: Long memory, level shifts, leverage effects, day-of-the-week seasonality, and macroeconomic announcements. *International Journal of forecasting 25*(2), 282–303.

Martens, M. and J. Zein (2004). Predicting financial volatility: High-frequency time-series forecasts vis-à-vis implied volatility. *Journal of Futures Markets: Futures, Options, and Other Derivative Products 24*(11), 1005–1028.

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

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Morgan, J. et al. (1996). Riskmetrics technical document.

Müller, U. A., M. M. Dacorogna, R. D. Davé, O. V. Pictet, R. B. Olsen, and J. R. Ward (1993). Fractals and intrinsic time: A challenge to econometricians. *Unpublished manuscript, Olsen & Associates, Zürich*.

Neely, C. J. (2009). Forecasting foreign exchange volatility: Why is implied volatility biased and inefficient? and does it matter? *Journal of International Financial Markets, Institutions and Money 19*(1), 188–205.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 347–370.

Oliveira, N., P. Cortez, and N. Areal (2017). The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications 73*, 125–144.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics 160*(1), 246–256.

Patton, A. J. and K. Sheppard (2009a). Evaluating volatility and correlation forecasts. In *Handbook of financial time series*, pp. 801–838. Springer.

Patton, A. J. and K. Sheppard (2009b). Optimal combinations of realised volatility estimators. *International Journal of Forecasting 25*(2), 218–238.

Patton, A. J. and K. Sheppard (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics 97*(3), 683–697.

Pesaran, M. H. and A. Timmermann (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics 137*(1), 134–161.

Pinho, D. M. (2020). Code for: "Forecast comparison of volatility models and their combinations for the FTSE 100: a tied race". `https://github.com/davidmpinho/volatility-models-comparison-thesis`.

Pong, S., M. B. Shackleton, S. J. Taylor, and X. Xu (2004). Forecasting currency volatility: A comparison of implied volatilities and ar (fi) ma models. *Journal of Banking & Finance 28*(10), 2541–2563.

Poon, S.-H. and C. Granger (2005). Practical issues in forecasting volatility. *Financial analysts journal*, 45–56.

Poon, S.-H. and C. W. Granger (2003). Forecasting volatility in financial markets: A review. *Journal of economic literature 41*(2), 478–539.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rapach, D. E. and J. K. Strauss (2008). Structural breaks and garch models of exchange rate volatility. *Journal of Applied Econometrics 23*(1), 65–90.

Rapach, D. E. and J. K. Strauss (2010). Bagging or combining (or both)? an analysis based on forecasting us employment growth. *Econometric Reviews 29*(5-6), 511–533.

Rapach, D. E., J. K. Strauss, and M. E. Wohar (2008). Chapter 10 forecasting stock return volatility in the presence of structural breaks. *Forecasting in the presence of structural breaks and model uncertainty*, 381–416.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies 23*(2), 821–862.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Schwert, G. W. (1989). Why does stock market volatility change over time? *The journal of finance 44*(5), 1115–1153.

Schwert, G. W. (1990). Stock volatility and the crash of '87. *The review of financial studies 3*(1), 77–102.

Siriopoulos, C. and A. Fassas (2008). The information content of vftse. *Available at SSRN 1307702*.

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

Stock, J. H. and M. W. Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics 30*(4), 481–493.

Taylor, S. J. (2008). *Modelling financial time series*. world scientific.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

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

Vivian, A. and M. E. Wohar (2012). Commodity volatility breaks. *Journal of International Financial Markets, Institutions and Money 22*(2), 395–422.

Wang, Y.-H., A. Keswani, and S. J. Taylor (2006). The relationships between sentiment, returns and volatility. *International Journal of Forecasting 22*(1), 109–123.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 1–25.

White, H. (2000). A reality check for data snooping. *Econometrica 68*(5), 1097–1126.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. Nelson Education.

Yang, K., F. Tian, L. Chen, and S. Li (2017). Realized volatility forecast of agricultural futures using the har models with bagging and combination approaches. *International Review of Economics & Finance 49*, 276–291.

Zakoian, J.-M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and control 18*(5), 931–955.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

Appendix A Data cleaning

All of the anomalies in the data were related to the intraday data, so FTSE100 and VFTSE indexes are only changed in one way: I exclude any days that are not included in the *clean* version of the intraday data.

To clean the intraday data, I firstly eliminate all observations that:

- Have the same day and hour (only the last of the repeated observations is not excluded.
- Are unordered in time.
- Are outside trading hours or trading days (includes holidays and weekends).

To find other mistakes in the data, I flag observations with:

- Large price changes (adjusted or not by the volatility of the respective day).
- Days with returns close or equal to zero.
- Large differences in time between intraday observations, or days.

I then sort anomalies by how large they are and investigate them individually by observing how they cluster, or how they related to news in those events. I stop investigating when there are many anomalies below a certain magnitude that have not been excluded.

A few issues stand out, either because the inclusion of observations became more subjective, or because they have a larger impact on the final results:

- On 2000-04-05, there was a technical problem with the London Stock Exchange because it opened and closed later than usual. The number of observations was similar, although my data set seems to miss the last hour of trading. I kept this observation because it is a typical trading day.
- From 2010-03-29 to 2010-03-31, which is after the timezone change in 2010-03-28, the trading schedule does not change to the one of British Summer Time. I could not find any warnings or news of this occurrence, but those days are very similar to a typical day, so I include them.
- There were a few other days not excluded because the London Stock Exchange had technical problems. I did not exclude them.
- I excluded two very large intraday returns on 2000-03-13, where the first one was of roughly +6%, followed almost immediately after by a second one of roughly -6% in a period where volatility did not seem to justify the magnitude of those returns, especially in such a short time. According to other (free) data sources, the price hit after the +6% was higher than the maximum price for that day.

- From 2001-09-13 to 2001-10-03 (after the terrorist attacks of 2001-09-11), there are 14 days even though other data sources indicate that the exchange was open.

As per Liu et al. (2015), after all these steps, I exclude observations that have less than 60% of the maximum number of 1-minute returns per day (i.e., exclude days with less than 306 observations), and then convert them to 5-minute returns.

Appendix B Summary statistics

Table 7: Summary statistics for volatility measures

Period	Name	Minimum	5^{th} perc.	Mean	95^{th} perc.	Maximum	SD	ACF(1)	ACF(2)	ACF(3)
2000-2004	RQ	$7.7e-11$	$6.6e-10$	$2.9e-07$	$5.1e-07$	$9.7e-05$	$3.4e-06$	0.01	0.02	0.03
	TRV	$3.5e-06$	$1.0e-05$	$6.9e-05$	$2.1e-04$	$2.6e-03$	$1.2e-04$	0.36	0.32	0.30
	RV^+	$3.6e-06$	$9.2e-06$	$5.9e-05$	$1.8e-04$	$1.9e-03$	$9.4e-05$	0.42	0.38	0.37
	RV^-	$3.5e-06$	$9.7e-06$	$5.2e-05$	$1.6e-04$	$8.3e-04$	$6.4e-05$	0.73	0.66	0.60
	RV	$7.0e-06$	$2.1e-05$	$1.2e-04$	$3.7e-04$	$3.2e-03$	$1.8e-04$	0.52	0.47	0.44
	r^2	$2.3e-10$	$4.4e-07$	$1.6e-04$	$7.1e-04$	$3.5e-03$	$3.4e-04$	0.21	0.35	0.25
2005-2006	RQ	$3.7e-11$	$1.6e-10$	$2.7e-08$	$1.0e-07$	$1.4e-06$	$1.2e-07$	0.12	0.17	0.49
	TRV	$1.8e-06$	$5.2e-06$	$2.3e-05$	$6.7e-05$	$3.0e-04$	$2.9e-05$	0.32	0.32	0.34
	RV^+	$2.5e-06$	$4.7e-06$	$1.9e-05$	$5.9e-05$	$2.2e-04$	$2.1e-05$	0.37	0.30	0.28
	RV^-	$3.3e-06$	$4.7e-06$	$1.5e-05$	$4.0e-05$	$1.4e-04$	$1.6e-05$	0.60	0.53	0.52
	RV	$5.3e-06$	$1.0e-05$	$3.8e-05$	$1.1e-04$	$4.4e-04$	$4.4e-05$	0.43	0.41	0.41
	r^2	$4.1e-10$	$1.7e-07$	$4.7e-05$	$1.9e-04$	$8.8e-04$	$9.1e-05$	0.23	0.18	0.37
2007-2010	RQ	$1.6e-10$	$1.2e-09$	$3.2e-06$	$2.1e-06$	$2.2e-03$	$7.0e-05$	0.07	0.03	0.01
	TRV	$5.9e-06$	$1.4e-05$	$1.3e-04$	$3.5e-04$	$8.4e-03$	$3.4e-04$	0.37	0.41	0.28
	RV^+	$4.6e-06$	$1.3e-05$	$1.1e-04$	$3.3e-04$	$2.6e-03$	$2.0e-04$	0.57	0.46	0.49
	RV^-	$4.4e-06$	$1.2e-05$	$9.4e-05$	$3.1e-04$	$2.5e-03$	$1.6e-04$	0.67	0.63	0.51
	RV	$1.2e-05$	$2.7e-05$	$2.2e-04$	$7.1e-04$	$1.1e-02$	$4.8e-04$	0.48	0.51	0.37
	r^2	$2.4e-12$	$4.3e-07$	$2.6e-04$	$1.0e-03$	$8.8e-03$	$7.2e-04$	0.21	0.26	0.30

Each statistic corresponds only to the data within the respective period. "perc." means "percentile"; SD is the standard deviation; and "ACF(t)" is the autocorrelation function with t lags. The names of the volatility measures are described in section 3.2.

Table 8: Average correlation by model category

	$ARMA^{RV}$	$ARMA^{r}$	EqualWeight	GARCH	Grouping	HAR	IV	$Naïve^{RV}$	$Naïve^{r}$	$Regularization^{MSE}$	$Regularization^{QLIKE}$	$Trimming^{MSE}$	$Trimming^{QLIKE}$
$ARMA^{RV}$	0.90	0.57	0.66	0.60	0.69	0.79	0.86	0.71	0.80	0.77	0.75	0.80	0.79
$ARMA^{r}$	0.57	0.83	0.65	0.62	0.66	0.78	0.68	0.80	0.70	0.71	0.73	0.74	0.71
EqualWeight	0.66	0.65	1.00	0.83	0.83	0.91	0.90	0.81	0.84	0.90	0.91	0.94	0.95
GARCH	0.60	0.62	0.83	0.75	0.74	0.82	0.80	0.75	0.76	0.78	0.80	0.84	0.84
Grouping	0.69	0.66	0.83	0.74	0.77	0.86	0.85	0.78	0.80	0.82	0.83	0.87	0.86
HAR	0.79	0.78	0.91	0.82	0.86	0.97	0.95	0.90	0.91	0.92	0.93	0.97	0.96
IV	0.86	0.68	0.90	0.80	0.85	0.95	1.00	0.85	0.92	0.93	0.93	0.96	0.96
$Naïve^{RV}$	0.71	0.80	0.81	0.75	0.78	0.90	0.85	0.89	0.84	0.84	0.86	0.89	0.87
$Naïve^{r}$	0.80	0.70	0.84	0.76	0.80	0.91	0.92	0.84	0.87	0.87	0.87	0.92	0.91
$Regularization^{MSE}$	0.77	0.71	0.90	0.78	0.82	0.92	0.93	0.84	0.87	0.90	0.91	0.93	0.93
$Regularization^{QLIKE}$	0.75	0.73	0.91	0.80	0.83	0.93	0.93	0.86	0.87	0.91	0.92	0.94	0.93
$Trimming^{MSE}$	0.80	0.74	0.94	0.84	0.87	0.97	0.96	0.89	0.92	0.93	0.94	0.99	0.99
$Trimming^{QLIKE}$	0.79	0.71	0.95	0.84	0.86	0.96	0.96	0.87	0.91	0.93	0.93	0.99	0.99

I first take all the 250 models that I tested and calculate their Pearson correlation coefficient, which creates a correlation matrix. I then average these coefficients groupwise (I include the diagonal of the matrix). When the groups are the same (e.g., $ARMA^{RV}$ with $ARMA^{RV}$), the average coefficient tells us how similar models of the same category are – that is why the coefficient for IV and EqualWeight is 1, as there is only one implied volatility and one "EqualWeight" model being tested.

Appendix C Empirical results

Table 9: Results (Part 1)

Model name	Model family	MSE_1	$QLIKE_1$	MSE_2	$QLIKE_2$	Avg_{MSE_2}	Avg_{QLIKE_2}	$\hat{p}_{MSE}$	$\hat{p}_{QLIKE}$
EqualWeight	EqualWeight	—	—	0.90	$1.0e+00$	0.90	1.00	1.00	1.00
Group($ARMA^{RV}$)	Grouping	—	—	0.99	$1.2e+00$	0.91	1.12	1.00	1.00
Group($ARMA^{RV}$, $ARMA^{r}$)	Grouping	—	—	0.97	$1.4e+00$	0.91	1.12	1.00	1.00
Group($ARMA^{RV}$, $ARMA^{r}$, IV)	Grouping	—	—	0.92	$1.3e+00$	0.91	1.12	1.00	1.00
Group($ARMA^{RV}$, IV)	Grouping	—	—	0.90	$1.2e+00$	0.91	1.12	1.00	1.00
Group($ARMA^{r}$)	Grouping	—	—	1.04	$1.7e+00$	0.91	1.12	0.98	1.00
Group($ARMA^{r}$, IV)	Grouping	—	—	0.95	$1.5e+00$	0.91	1.12	1.00	1.00
Group(GARCH)	Grouping	—	—	0.86	$9.8e-01$	0.91	1.12	1.00	1.00
Group(GARCH, $ARMA^{RV}$)	Grouping	—	—	0.86	$9.9e-01$	0.91	1.12	1.00	1.00
Group(GARCH, $ARMA^{RV}$, $ARMA^{r}$)	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, $ARMA^{RV}$, $ARMA^{r}$, IV)	Grouping	—	—	0.89	$1.2e+00$	0.91	1.12	1.00	1.00
Group(GARCH, $ARMA^{RV}$, IV)	Grouping	—	—	0.87	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, $ARMA^{r}$)	Grouping	—	—	0.91	$1.2e+00$	0.91	1.12	1.00	1.00
Group(GARCH, $ARMA^{r}$, IV)	Grouping	—	—	0.90	$1.3e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR)	Grouping	—	—	0.88	$9.1e-01$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $ARMA^{RV}$)	Grouping	—	—	0.88	$9.5e-01$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $ARMA^{RV}$, $ARMA^{r}$)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $ARMA^{RV}$, $ARMA^{r}$, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $ARMA^{RV}$, IV)	Grouping	—	—	0.87	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $ARMA^{r}$)	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $ARMA^{r}$, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, IV)	Grouping	—	—	0.88	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $Naïve^{RV}$)	Grouping	—	—	0.88	$9.3e-01$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $Naïve^{RV}$, $ARMA^{RV}$)	Grouping	—	—	0.88	$9.6e-01$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $Naïve^{RV}$, $ARMA^{RV}$, $ARMA^{r}$)	Grouping	—	—	0.89	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $Naïve^{RV}$, $ARMA^{RV}$, $ARMA^{r}$, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $Naïve^{RV}$, $ARMA^{RV}$, IV)	Grouping	—	—	0.88	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $Naïve^{RV}$, $ARMA^{r}$)	Grouping	—	—	0.89	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, $Naïve^{RV}$, $ARMA^{r}$, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00

The "model family" is what I often refer to as the "model category" of each model. "MSE_1"/"$QLIKE_1$" is the model's loss in the period of 2005-2006 relative to the loss of the `HAR-RV` in 2005-2006. "MSE_2"/"$QLIKE_2$" is the model's loss in the period of 2007-2010 relative to the loss of the `HAR-RV` in 2005-2006. "Avg_{MSE_2}"/"Avg_{MSE_2}" is the average relative loss of the MODEL FAMILY. "$\hat{p}_{MSE}$"/"$\hat{p}_{QLIKE}$" are the MCS p-values for the model ("Model name") using the MSE and QLIKE loss functions, respectively. Empty values ("—") in those p-values signify that the model was excluded from the set of superior models (please refer to section 4.4 for more details on the MCS test).

Table 10: Results (Part 2)

Model name	Model family	MSE_1	QLIKE_1	MSE_2	QLIKE_2	$\text{Avg}_{\text{MSE}_2}$	$\text{Avg}_{\text{QLIKE}_2}$	$\hat{\text{p}}_{\text{MSE}}$	$\hat{\text{p}}_{\text{QLIKE}}$
Group(GARCH, HAR, Naïve^{RV}, IV)	Grouping	—	—	0.88	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{RV}, Naïve^{r})	Grouping	—	—	0.89	$9.7e-01$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{RV}, Naïve^{r}, ARMA^{RV})	Grouping	—	—	0.89	$9.8e-01$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{RV}, Naïve^{r}, ARMA^{RV}, ARMA^{r})	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{RV}, Naïve^{r}, ARMA^{RV}, ARMA^{r}, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{RV}, Naïve^{r}, ARMA^{RV}, IV)	Grouping	—	—	0.88	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{RV}, Naïve^{r}, ARMA^{r})	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{RV}, Naïve^{r}, ARMA^{r}, IV)	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{RV}, Naïve^{r}, IV)	Grouping	—	—	0.89	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{r})	Grouping	—	—	0.90	$9.8e-01$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{r}, ARMA^{RV})	Grouping	—	—	0.89	$9.8e-01$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{r}, ARMA^{RV}, ARMA^{r})	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{r}, ARMA^{RV}, ARMA^{r}, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{r}, ARMA^{RV}, IV)	Grouping	—	—	0.88	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{r}, ARMA^{r})	Grouping	—	—	0.91	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{r}, ARMA^{r}, IV)	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, HAR, Naïve^{r}, IV)	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, IV)	Grouping	—	—	0.90	$1.2e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV})	Grouping	—	—	0.86	$9.6e-01$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, ARMA^{RV})	Grouping	—	—	0.88	$9.9e-01$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, ARMA^{RV}, ARMA^{r})	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, ARMA^{RV}, ARMA^{r}, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, ARMA^{RV}, IV)	Grouping	—	—	0.87	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, ARMA^{r})	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, ARMA^{r}, IV)	Grouping	—	—	0.89	$1.2e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, IV)	Grouping	—	—	0.88	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, Naïve^{r})	Grouping	—	—	0.89	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, Naïve^{r}, ARMA^{RV})	Grouping	—	—	0.89	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, Naïve^{r}, ARMA^{RV}, ARMA^{r})	Grouping	—	—	0.91	$1.1e+00$	0.91	1.12	1.00	1.00

The "model family" is what I often refer to as the "model category" of each model. "MSE_1"/"QLIKE_1" is the model's loss in the period of 2005-2006 relative to the loss of the `HAR-RV` in 2005-2006. "MSE_2"/"QLIKE_2" is the model's loss in the period of 2007-2010 relative to the loss of the `HAR-RV` in 2005-2006. "$\text{Avg}_{\text{MSE}_2}$"/"$\text{Avg}_{\text{MSE}_2}$" is the average relative loss of the MODEL FAMILY. "$\hat{\text{p}}_{\text{MSE}}$"/"$\hat{\text{p}}_{\text{QLIKE}}$" are the MCS p-values for the model ("Model name") using the MSE and QLIKE loss functions, respectively. Empty values ("—") in those p-values signify that the model was excluded from the set of superior models (please refer to section 4.4 for more details on the MCS test).

Table 11: Results (Part 3)

Model name	Model family	MSE_1	$QLIKE_1$	MSE_2	$QLIKE_2$	Avg_{MSE_2}	Avg_{QLIKE_2}	$\hat{p}_{MSE}$	$\hat{p}_{QLIKE}$
Group(GARCH, Naïve^{RV}, Naïve^{r}, ARMA^{RV}, ARMA^{r}, IV)	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, Naïve^{r}, ARMA^{RV}, IV)	Grouping	—	—	0.88	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, Naïve^{r}, ARMA^{r})	Grouping	—	—	0.91	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, Naïve^{r}, ARMA^{r}, IV)	Grouping	—	—	0.90	$1.2e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{RV}, Naïve^{r}, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{r})	Grouping	—	—	0.92	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{r}, ARMA^{RV})	Grouping	—	—	0.89	$1.0e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{r}, ARMA^{RV}, ARMA^{r})	Grouping	—	—	0.91	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{r}, ARMA^{RV}, ARMA^{r}, IV)	Grouping	—	—	0.90	$1.2e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{r}, ARMA^{RV}, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{r}, ARMA^{r})	Grouping	—	—	0.94	$1.2e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{r}, ARMA^{r}, IV)	Grouping	—	—	0.92	$1.2e+00$	0.91	1.12	1.00	1.00
Group(GARCH, Naïve^{r}, IV)	Grouping	—	—	0.91	$1.2e+00$	0.91	1.12	1.00	1.00
Group(HAR)	Grouping	—	—	0.94	$9.2e-01$	0.91	1.12	1.00	1.00
Group(HAR, ARMA^{RV})	Grouping	—	—	0.92	$1.0e+00$	0.91	1.12	1.00	1.00
Group(HAR, ARMA^{RV}, ARMA^{r})	Grouping	—	—	0.93	$1.2e+00$	0.91	1.12	1.00	1.00
Group(HAR, ARMA^{RV}, ARMA^{r}, IV)	Grouping	—	—	0.90	$1.2e+00$	0.91	1.12	1.00	1.00
Group(HAR, ARMA^{RV}, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(HAR, ARMA^{r})	Grouping	—	—	0.94	$1.2e+00$	0.91	1.12	1.00	1.00
Group(HAR, ARMA^{r}, IV)	Grouping	—	—	0.91	$1.2e+00$	0.91	1.12	1.00	1.00
Group(HAR, IV)	Grouping	—	—	0.91	$1.1e+00$	0.91	1.12	1.00	1.00
Group(HAR, Naïve^{RV})	Grouping	—	—	0.91	$9.7e-01$	0.91	1.12	1.00	1.00
Group(HAR, Naïve^{RV}, ARMA^{RV})	Grouping	—	—	0.91	$1.0e+00$	0.91	1.12	1.00	1.00
Group(HAR, Naïve^{RV}, ARMA^{RV}, ARMA^{r})	Grouping	—	—	0.92	$1.1e+00$	0.91	1.12	1.00	1.00
Group(HAR, Naïve^{RV}, ARMA^{RV}, ARMA^{r}, IV)	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(HAR, Naïve^{RV}, ARMA^{RV}, IV)	Grouping	—	—	0.89	$1.1e+00$	0.91	1.12	1.00	1.00
Group(HAR, Naïve^{RV}, ARMA^{r})	Grouping	—	—	0.92	$1.1e+00$	0.91	1.12	1.00	1.00
Group(HAR, Naïve^{RV}, ARMA^{r}, IV)	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00
Group(HAR, Naïve^{RV}, IV)	Grouping	—	—	0.90	$1.1e+00$	0.91	1.12	1.00	1.00

The "model family" is what I often refer to as the "model category" of each model. "MSE_1"/"$QLIKE_1$" is the model's loss in the period of 2005-2006 relative to the loss of the `HAR-RV` in 2005-2006. "MSE_2"/"$QLIKE_2$" is the model's loss in the period of 2007-2010 relative to the loss of the `HAR-RV` in 2005-2006. "Avg_{MSE_2}"/"Avg_{MSE_2}" is the average relative loss of the MODEL FAMILY. "$\hat{p}_{MSE}$"/"$\hat{p}_{QLIKE}$" are the MCS p-values for the model ("Model name") using the MSE and QLIKE loss functions, respectively. Empty values ("—") in those p-values signify that the model was excluded from the set of superior models (please refer to section 4.4 for more details on the MCS test).

Table 12: Results (Part 4)

Model name	Model family	MSE_1	QLIKE_1	MSE_2	QLIKE_2	$\text{Avg}_{\text{MSE}_2}$	$\text{Avg}_{\text{QLIKE}_2}$	$\hat{\text{p}}_{\text{MSE}}$	$\hat{\text{p}}_{\text{QLIKE}}$
Group(HAR, Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$)	Grouping	—	—	0.92	1.0e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$)	Grouping	—	—	0.91	1.0e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$, ARMA$^{\text{r}}$)	Grouping	—	—	0.92	1.1e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$, ARMA$^{\text{r}}$, IV)	Grouping	—	—	0.91	1.1e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$, IV)	Grouping	—	—	0.90	1.1e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, ARMA$^{\text{r}}$)	Grouping	—	—	0.93	1.1e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, ARMA$^{\text{r}}$, IV)	Grouping	—	—	0.91	1.1e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, IV)	Grouping	—	—	0.90	1.1e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{r}}$)	Grouping	—	—	0.95	1.0e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$)	Grouping	—	—	0.92	1.0e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$, ARMA$^{\text{r}}$)	Grouping	—	—	0.93	1.1e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$, ARMA$^{\text{r}}$, IV)	Grouping	—	—	0.91	1.2e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$, IV)	Grouping	—	—	0.90	1.1e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{r}}$, ARMA$^{\text{r}}$)	Grouping	—	—	0.95	1.2e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{r}}$, ARMA$^{\text{r}}$, IV)	Grouping	—	—	0.92	1.2e + 00	0.91	1.12	1.00	1.00
Group(HAR, Naïve$^{\text{r}}$, IV)	Grouping	—	—	0.92	1.1e + 00	0.91	1.12	1.00	1.00
Group(IV)	Grouping	—	—	0.98	1.6e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$)	Grouping	—	—	0.91	1.1e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, ARMA$^{\text{RV}}$)	Grouping	—	—	0.94	1.1e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, ARMA$^{\text{RV}}$, ARMA$^{\text{r}}$)	Grouping	—	—	0.94	1.2e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, ARMA$^{\text{RV}}$, ARMA$^{\text{r}}$, IV)	Grouping	—	—	0.91	1.2e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, ARMA$^{\text{RV}}$, IV)	Grouping	—	—	0.90	1.1e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, ARMA$^{\text{r}}$)	Grouping	—	—	0.94	1.2e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, ARMA$^{\text{r}}$, IV)	Grouping	—	—	0.92	1.3e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, IV)	Grouping	—	—	0.91	1.2e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$)	Grouping	—	—	0.94	1.1e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$)	Grouping	—	—	0.93	1.1e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$, ARMA$^{\text{r}}$)	Grouping	—	—	0.94	1.2e + 00	0.91	1.12	1.00	1.00
Group(Naïve$^{\text{RV}}$, Naïve$^{\text{r}}$, ARMA$^{\text{RV}}$, ARMA$^{\text{r}}$, IV)	Grouping	—	—	0.92	1.2e + 00	0.91	1.12	1.00	1.00

The "model family" is what I often refer to as the "model category" of each model. "MSE_1"/"QLIKE_1" is the model's loss in the period of 2005-2006 relative to the loss of the `HAR-RV` in 2005-2006. "MSE_2"/"QLIKE_2" is the model's loss in the period of 2007-2010 relative to the loss of the `HAR-RV` in 2005-2006. "$\text{Avg}_{\text{MSE}_2}$"/"$\text{Avg}_{\text{MSE}_2}$" is the average relative loss of the MODEL FAMILY. "$\hat{\text{p}}_{\text{MSE}}$"/"$\hat{\text{p}}_{\text{QLIKE}}$" are the MCS p-values for the model ("Model name") using the MSE and QLIKE loss functions, respectively. Empty values ("—") in those p-values signify that the model was excluded from the set of superior models (please refer to section 4.4 for more details on the MCS test).

Table 13: Results (Part 5)

Model name	Model family	MSE_1	$QLIKE_1$	MSE_2	$QLIKE_2$	Avg_{MSE_2}	Avg_{QLIKE_2}	$\hat{p}_{MSE}$	$\hat{p}_{QLIKE}$
Group(Naïve^{RV}, Naïve^{r}, ARMA^{RV}, IV)	Grouping	—	—	0.90	1.1e + 00	0.91	1.12	1.00	1.00
Group(Naïve^{RV}, Naïve^{r}, ARMA^{r})	Grouping	—	—	0.95	1.2e + 00	0.91	1.12	1.00	1.00
Group(Naïve^{RV}, Naïve^{r}, ARMA^{r}, IV)	Grouping	—	—	0.93	1.2e + 00	0.91	1.12	1.00	1.00
Group(Naïve^{RV}, Naïve^{r}, IV)	Grouping	—	—	0.92	1.2e + 00	0.91	1.12	1.00	1.00
Group(Naïve^{r})	Grouping	—	—	1.07	1.3e + 00	0.91	1.12	0.75	1.00
Group(Naïve^{r}, ARMA^{RV})	Grouping	—	—	0.95	1.1e + 00	0.91	1.12	1.00	1.00
Group(Naïve^{r}, ARMA^{RV}, ARMA^{r})	Grouping	—	—	0.96	1.3e + 00	0.91	1.12	1.00	1.00
Group(Naïve^{r}, ARMA^{r})	Grouping	—	—	1.01	1.4e + 00	0.91	1.12	1.00	1.00
Group(Naïve^{r}, ARMA^{r}, IV)	Grouping	—	—	0.95	1.4e + 00	0.91	1.12	1.00	1.00
Group(Naïve^{r}, IV)	Grouping	—	—	0.96	1.3e + 00	0.91	1.12	1.00	1.00
ElasticNet^{MSE}	Regularization^{MSE}	—	—	2.29	1.5e + 00	1.90	2.03	0.85	1.00
Lasso^{MSE}	Regularization^{MSE}	—	—	2.28	3.6e + 00	1.90	2.03	0.85	0.95
Ridge^{MSE}	Regularization^{MSE}	—	—	1.13	1.0e + 00	1.90	2.03	0.95	1.00
ElasticNet^{QLIKE}	Regularization^{QLIKE}	—	—	1.21	2.1e + 00	1.21	2.97	0.71	1.00
Lasso^{QLIKE}	Regularization^{QLIKE}	—	—	1.35	2.2e + 00	1.21	2.97	0.63	1.00
Ridge^{QLIKE}	Regularization^{QLIKE}	—	—	1.07	4.6e + 00	1.21	2.97	0.86	0.90
Trimming^{MSE}(1.1)	Trimming^{MSE}	—	—	0.87	9.3e − 01	0.89	0.98	1.00	1.00
Trimming^{MSE}(1.2)	Trimming^{MSE}	—	—	0.89	9.0e − 01	0.89	0.98	1.00	1.00
Trimming^{MSE}(1.3)	Trimming^{MSE}	—	—	0.89	9.1e − 01	0.89	0.98	1.00	1.00
Trimming^{MSE}(1.4)	Trimming^{MSE}	—	—	0.89	9.2e − 01	0.89	0.98	1.00	1.00
Trimming^{MSE}(1.5)	Trimming^{MSE}	—	—	0.88	9.2e − 01	0.89	0.98	1.00	1.00
Trimming^{MSE}(1.6)	Trimming^{MSE}	—	—	0.87	9.4e − 01	0.89	0.98	1.00	1.00
Trimming^{MSE}(1.7)	Trimming^{MSE}	—	—	0.88	9.8e − 01	0.89	0.98	1.00	1.00
Trimming^{MSE}(1.8)	Trimming^{MSE}	—	—	0.88	1.0e + 00	0.89	0.98	1.00	1.00
Trimming^{MSE}(1.9)	Trimming^{MSE}	—	—	0.88	1.0e + 00	0.89	0.98	1.00	1.00
Trimming^{MSE}(1)	Trimming^{MSE}	—	—	0.89	1.1e + 00	0.89	0.98	1.00	1.00
Trimming^{MSE}(2.1)	Trimming^{MSE}	—	—	0.89	1.0e + 00	0.89	0.98	1.00	1.00
Trimming^{MSE}(2.2)	Trimming^{MSE}	—	—	0.89	1.0e + 00	0.89	0.98	1.00	1.00
Trimming^{MSE}(2.3)	Trimming^{MSE}	—	—	0.89	1.0e + 00	0.89	0.98	1.00	1.00

The "model family" is what I often refer to as the "model category" of each model. "MSE_1"/"$QLIKE_1$" is the model's loss in the period of 2005-2006 relative to the loss of the HAR-RV in 2005-2006. "MSE_2"/"$QLIKE_2$" is the model's loss in the period of 2007-2010 relative to the loss of the HAR-RV in 2005-2006. "Avg_{MSE_2}"/"Avg_{MSE_2}" is the average relative loss of the MODEL FAMILY. "$\hat{p}_{MSE}$"/"$\hat{p}_{QLIKE}$" are the MCS p-values for the model ("Model name") using the MSE and QLIKE loss functions, respectively. Empty values ("—") in those p-values signify that the model was excluded from the set of superior models (please refer to section 4.4 for more details on the MCS test).

Table 14: Results (Part 6)

Model name	Model family	MSE_1	$QLIKE_1$	MSE_2	$QLIKE_2$	Avg_{MSE_2}	Avg_{QLIKE_2}	$\hat{p}_{MSE}$	$\hat{p}_{QLIKE}$
$\texttt{Trimming}^{MSE}(2.4)$	Trimming^{MSE}	—	—	0.89	1.0e + 00	0.89	0.98	1.00	1.00
$\texttt{Trimming}^{MSE}(2.5)$	Trimming^{MSE}	—	—	0.90	1.0e + 00	0.89	0.98	1.00	1.00
$\texttt{Trimming}^{MSE}(2.6)$	Trimming^{MSE}	—	—	0.90	1.0e + 00	0.89	0.98	1.00	1.00
$\texttt{Trimming}^{MSE}(2.7)$	Trimming^{MSE}	—	—	0.90	1.0e + 00	0.89	0.98	1.00	1.00
$\texttt{Trimming}^{MSE}(2.8)$	Trimming^{MSE}	—	—	0.90	1.0e + 00	0.89	0.98	1.00	1.00
$\texttt{Trimming}^{MSE}(2.9)$	Trimming^{MSE}	—	—	0.90	1.0e + 00	0.89	0.98	1.00	1.00
$\texttt{Trimming}^{MSE}(2)$	Trimming^{MSE}	—	—	0.88	1.0e + 00	0.89	0.98	1.00	1.00
$\texttt{Trimming}^{MSE}(3)$	Trimming^{MSE}	—	—	0.90	1.0e + 00	0.89	0.98	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1.1)$	Trimming^{QLIKE}	—	—	0.90	8.9e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1.2)$	Trimming^{QLIKE}	—	—	0.88	8.8e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1.3)$	Trimming^{QLIKE}	—	—	0.88	9.0e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1.4)$	Trimming^{QLIKE}	—	—	0.89	9.1e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1.5)$	Trimming^{QLIKE}	—	—	0.88	9.2e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1.6)$	Trimming^{QLIKE}	—	—	0.89	9.2e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1.7)$	Trimming^{QLIKE}	—	—	0.89	9.2e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1.8)$	Trimming^{QLIKE}	—	—	0.89	9.2e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1.9)$	Trimming^{QLIKE}	—	—	0.88	9.3e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(1)$	Trimming^{QLIKE}	—	—	0.99	9.4e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2.1)$	Trimming^{QLIKE}	—	—	0.89	9.3e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2.2)$	Trimming^{QLIKE}	—	—	0.88	9.3e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2.3)$	Trimming^{QLIKE}	—	—	0.88	9.3e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2.4)$	Trimming^{QLIKE}	—	—	0.88	9.3e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2.5)$	Trimming^{QLIKE}	—	—	0.88	9.4e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2.6)$	Trimming^{QLIKE}	—	—	0.88	9.4e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2.7)$	Trimming^{QLIKE}	—	—	0.88	9.4e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2.8)$	Trimming^{QLIKE}	—	—	0.88	9.5e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2.9)$	Trimming^{QLIKE}	—	—	0.88	9.5e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(2)$	Trimming^{QLIKE}	—	—	0.89	9.3e − 01	0.89	0.93	1.00	1.00
$\texttt{Trimming}^{QLIKE}(3)$	Trimming^{QLIKE}	—	—	0.88	9.5e − 01	0.89	0.93	1.00	1.00

The "model family" is what I often refer to as the "model category" of each model. "MSE_1"/"$QLIKE_1$" is the model's loss in the period of 2005-2006 relative to the loss of the `HAR-RV` in 2005-2006. "MSE_2"/"$QLIKE_2$" is the model's loss in the period of 2007-2010 relative to the loss of the `HAR-RV` in 2005-2006. "Avg_{MSE_2}"/"Avg_{MSE_2}" is the average relative loss of the MODEL FAMILY. "$\hat{p}_{MSE}$"/"$\hat{p}_{QLIKE}$" are the MCS p-values for the model ("Model name") using the MSE and QLIKE loss functions, respectively. Empty values ("—") in those p-values signify that the model was excluded from the set of superior models (please refer to section 4.4 for more details on the MCS test).

Table 15: Results (Part 7)

Model name	Model family	MSE_1	$QLIKE_1$	MSE_2	$QLIKE_2$	Avg_{MSE_2}	Avg_{QLIKE_2}	$\hat{p}_{MSE}$	$\hat{p}_{QLIKE}$
$\texttt{ARMA}^{RV}(0,1)$	ARMA^{RV}	2.96	2.50	1.13	$2.5e+00$	1.07	1.63	0.77	1.00
$\texttt{ARMA}^{RV}(1,0)$	ARMA^{RV}	1.82	1.81	1.13	$1.4e+00$	1.07	1.63	0.96	1.00
$\texttt{ARMA}^{RV}(1,1)$	ARMA^{RV}	1.02	1.06	0.94	$9.9e-01$	1.07	1.63	1.00	1.00
$\texttt{ARMA}^{r}(0,1)$	ARMA^{r}	6.97	3.64	1.25	$4.0e+00$	1.15	2.74	0.75	0.88
$\texttt{ARMA}^{r}(1,0)$	ARMA^{r}	5.75	3.35	1.21	$3.0e+00$	1.15	2.74	0.69	1.00
$\texttt{ARMA}^{r}(1,1)$	ARMA^{r}	1.47	1.42	1.00	$1.2e+00$	1.15	2.74	1.00	1.00
C-GARCH(1,1)	GARCH	1.16	1.00	0.96	$1.2e+00$	0.94	1.33	1.00	1.00
E-GARCH(1,1)	GARCH	1.07	1.06	0.84	$1.1e+00$	0.94	1.33	1.00	1.00
F-GARCH(1,1)	GARCH	1.09	0.97	0.86	$1.1e+00$	0.94	1.33	1.00	1.00
AP-ARCH(1,1)	GARCH	1.10	0.96	0.86	$1.1e+00$	0.94	1.33	1.00	1.00
AV-GARCH(1,1)	GARCH	1.05	1.02	0.85	$1.1e+00$	0.94	1.33	1.00	1.00
GARCH(1,1)	GARCH	1.26	1.05	1.00	$1.2e+00$	0.94	1.33	1.00	1.00
GJR-GARCH(1,1)	GARCH	1.22	0.97	0.87	$1.1e+00$	0.94	1.33	1.00	1.00
NA-ARCH(1,1)	GARCH	1.05	1.01	0.85	$1.1e+00$	0.94	1.33	1.00	1.00
N-ARCH(1,1)	GARCH	1.24	1.05	0.98	$1.2e+00$	0.94	1.33	1.00	1.00
T-GARCH(1,1)	GARCH	1.07	0.96	0.86	$1.0e+00$	0.94	1.33	1.00	1.00
I-GARCH(1,1)	GARCH	1.33	1.05	1.01	$1.2e+00$	0.94	1.33	1.00	1.00
$\texttt{R-GARCH}^{r}(1,1)$	GARCH	1.10	0.96	0.98	$1.1e+00$	0.94	1.33	1.00	1.00
$\texttt{R-GARCH}^{RV}(1,1)$	GARCH	0.93	0.98	0.92	$1.1e+00$	0.94	1.33	1.00	1.00
ARCH(1)	GARCH	4.83	3.11	1.39	$4.2e+00$	0.94	1.33	0.69	0.85
Q-HAR-C-J	HAR	1.00	0.90	1.00	$9.3e-01$	1.01	0.99	1.00	1.00
HAR-C-J	HAR	0.97	0.94	0.97	$9.1e-01$	1.01	0.99	1.00	1.00
Q-HAR-C	HAR	1.17	0.96	1.07	$1.1e+00$	1.01	0.99	0.98	1.00
HAR-C	HAR	1.09	1.11	1.05	$1.2e+00$	1.01	0.99	0.99	1.00
QS-HAR-C-J	HAR	1.03	0.88	0.99	$9.3e-01$	1.01	0.99	1.00	1.00
S-HAR-C-J	HAR	0.98	0.93	0.97	$9.6e-01$	1.01	0.99	1.00	1.00
QS-HAR-C	HAR	1.16	0.89	1.00	$9.9e-01$	1.01	0.99	1.00	1.00
S-HAR-C	HAR	1.06	1.04	1.13	$1.1e+00$	1.01	0.99	0.97	1.00
Q-HAR-RV	HAR	1.08	0.89	1.01	$9.7e-01$	1.01	0.99	1.00	1.00

The "model family" is what I often refer to as the "model category" of each model. "MSE_1"/"$QLIKE_1$" is the model's loss in the period of 2005-2006 relative to the loss of the HAR-RV in 2005-2006. "MSE_2"/"$QLIKE_2$" is the model's loss in the period of 2007-2010 relative to the loss of the HAR-RV in 2005-2006. "Avg_{MSE_2}"/"Avg_{MSE_2}" is the average relative loss of the MODEL FAMILY. "$\hat{p}_{MSE}$"/"$\hat{p}_{QLIKE}$" are the MCS p-values for the model ("Model name") using the MSE and QLIKE loss functions, respectively. Empty values ("—") in those p-values signify that the model was excluded from the set of superior models (please refer to section 4.4 for more details on the MCS test).

Table 16: Results (Part 8)

Model name	Model family	MSE_1	$QLIKE_1$	MSE_2	$QLIKE_2$	Avg_{MSE_2}	Avg_{QLIKE_2}	$\hat{p}_{MSE}$	$\hat{p}_{QLIKE}$
QS-HAR-RV	HAR	1.10	0.88	0.97	$9.5e-01$	1.01	0.99	1.00	1.00
HAR-RV	HAR	1.00	1.00	1.00	$1.0e+00$	1.01	0.99	1.00	1.00
HAR-RVlog	HAR	0.93	0.88	0.90	$9.7e-01$	1.01	0.99	1.00	1.00
S-HAR-RV	HAR	1.00	0.98	1.07	$9.8e-01$	1.01	0.99	0.99	1.00
$RandWalk^r$	IV	1.65	1.56	0.98	$1.6e+00$	0.98	1.60	1.00	1.00
$ExpSmooth^{RV}(0.95)$	$Naïve^{RV}$	1.08	1.07	1.03	$1.4e+00$	1.16	1.99	0.97	1.00
$ExpSmooth^{RV}(0.9)$	$Naïve^{RV}$	0.99	1.00	0.95	$1.2e+00$	1.16	1.99	1.00	1.00
$ExpSmooth^{RV}(0.8)$	$Naïve^{RV}$	0.95	0.96	0.91	$1.0e+00$	1.16	1.99	1.00	1.00
$ExpSmooth^{RV}(0.7)$	$Naïve^{RV}$	0.96	0.96	0.91	$9.7e-01$	1.16	1.99	1.00	1.00
$ExpSmooth^{RV}(0.6)$	$Naïve^{RV}$	0.99	0.97	0.93	$9.6e-01$	1.16	1.99	1.00	1.00
$ExpSmooth^{RV}(0.5)$	$Naïve^{RV}$	1.04	0.98	0.97	$9.8e-01$	1.16	1.99	1.00	1.00
$ExpSmooth^{RV}(0.4)$	$Naïve^{RV}$	1.10	0.99	1.02	$1.0e+00$	1.16	1.99	1.00	1.00
$ExpSmooth^{RV}(0.3)$	$Naïve^{RV}$	1.17	1.01	1.09	$1.1e+00$	1.16	1.99	0.97	1.00
$ExpSmooth^{RV}(0.2)$	$Naïve^{RV}$	1.26	1.04	1.18	$1.1e+00$	1.16	1.99	0.95	1.00
$ExpSmooth^{RV}(0.1)$	$Naïve^{RV}$	1.37	1.08	1.28	$1.2e+00$	1.16	1.99	0.92	1.00
$ExpSmooth^{RV}(0.05)$	$Naïve^{RV}$	1.44	1.11	1.34	$1.2e+00$	1.16	1.99	0.90	1.00
$RandWalk^{RV}$	$Naïve^{RV}$	1.51	1.16	1.41	$1.3e+00$	1.16	1.99	0.89	1.00
$RollAvg^{RV}(1008)$	$Naïve^{RV}$	4.12	2.87	1.44	$6.7e+00$	1.16	1.99	0.61	0.33
$RollAvg^{RV}(126)$	$Naïve^{RV}$	1.34	1.37	1.28	$2.4e+00$	1.16	1.99	0.61	1.00
$RollAvg^{RV}(21)$	$Naïve^{RV}$	1.08	1.09	1.05	$1.4e+00$	1.16	1.99	0.94	1.00
$RollAvg^{RV}(252)$	$Naïve^{RV}$	1.32	1.38	1.35	$3.2e+00$	1.16	1.99	0.49	1.00
$HistVol^{RV}$	$Naïve^{RV}$	4.69	3.12	1.42	$5.0e+00$	1.16	1.99	0.67	0.79
$RollAvg^{RV}(5)$	$Naïve^{RV}$	0.99	1.06	0.95	$1.0e+00$	1.16	1.99	1.00	1.00
$RollAvg^{RV}(504)$	$Naïve^{RV}$	1.49	1.62	1.43	$4.7e+00$	1.16	1.99	0.50	0.73
$RollAvg^{RV}(63)$	$Naïve^{RV}$	1.28	1.25	1.21	$1.9e+00$	1.16	1.99	0.56	1.00
$ExpSmooth^r(0.95)$	$Naïve^r$	1.17	1.03	1.04	$1.4e+00$	1.56	7899.89	0.99	1.00
$ExpSmooth^r(0.9)$	$Naïve^r$	1.18	0.96	1.01	$1.2e+00$	1.56	7899.89	1.00	1.00
$ExpSmooth^r(0.8)$	$Naïve^r$	1.38	1.05	1.05	$1.3e+00$	1.56	7899.89	0.97	1.00
$ExpSmooth^r(0.7)$	$Naïve^r$	1.61	1.35	1.15	$1.7e+00$	1.56	7899.89	0.49	1.00

The "model family" is what I often refer to as the "model category" of each model. "MSE_1"/"$QLIKE_1$" is the model's loss in the period of 2005-2006 relative to the loss of the HAR-RV in 2005-2006. "MSE_2"/"$QLIKE_2$" is the model's loss in the period of 2007-2010 relative to the loss of the HAR-RV in 2005-2006. "Avg_{MSE_2}"/"Avg_{MSE_2}" is the average relative loss of the MODEL FAMILY. "$\hat{p}_{MSE}$"/"$\hat{p}_{QLIKE}$" are the MCS p-values for the model ("Model name") using the MSE and QLIKE loss functions, respectively. Empty values ("—") in those p-values signify that the model was excluded from the set of superior models (please refer to section 4.4 for more details on the MCS test).

Table 17: Results (Part 9)

Model name	Model family	MSE_1	$QLIKE_1$	MSE_2	$QLIKE_2$	Avg_{MSE_2}	Avg_{QLIKE_2}	$\hat{p}_{MSE}$	$\hat{p}_{QLIKE}$
`ExpSmooth`^{r}(0.6)	Naïve^{r}	1.86	1.84	1.28	$2.5e+00$	1.56	7899.89	0.37	1.00
`ExpSmooth`^{r}(0.5)	Naïve^{r}	2.15	2.59	1.44	$3.7e+00$	1.56	7899.89	0.35	0.81
`ExpSmooth`^{r}(0.4)	Naïve^{r}	2.48	3.73	1.64	$6.1e+00$	1.56	7899.89	0.34	—
`ExpSmooth`^{r}(0.3)	Naïve^{r}	2.89	5.60	1.87	$1.1e+01$	1.56	7899.89	0.34	—
`ExpSmooth`^{r}(0.2)	Naïve^{r}	3.37	9.29	2.16	$2.8e+01$	1.56	7899.89	0.34	0.24
`ExpSmooth`^{r}(0.1)	Naïve^{r}	3.96	19.72	2.51	$1.1e+02$	1.56	7899.89	0.34	0.45
`ExpSmooth`^{r}(0.05)	Naïve^{r}	4.30	37.31	2.71	$2.7e+02$	1.56	7899.89	0.34	0.51
`RandWalk`^{r}	Naïve^{r}	4.67	1093.65	2.94	$1.6e+05$	1.56	7899.89	0.34	—
`RollAvg`^{r}(5)	Naïve^{r}	1.78	2.05	1.16	$2.5e+00$	1.56	7899.89	0.62	1.00
`RollAvg`^{r}(126)	Naïve^{r}	1.54	1.40	1.31	$2.1e+00$	1.56	7899.89	0.36	1.00
`RollAvg`^{r}(21)	Naïve^{r}	1.32	1.04	1.08	$1.5e+00$	1.56	7899.89	0.98	1.00
`RollAvg`^{r}(63)	Naïve^{r}	1.49	1.23	1.27	$1.8e+00$	1.56	7899.89	0.43	1.00
`RollAvg`^{r}(252)	Naïve^{r}	1.36	1.37	1.38	$2.8e+00$	1.56	7899.89	0.34	1.00
`RollAvg`^{r}(504)	Naïve^{r}	1.65	1.65	1.44	$4.0e+00$	1.56	7899.89	0.39	0.86
`RollAvg`^{r}(1008)	Naïve^{r}	7.82	3.69	1.43	$5.4e+00$	1.56	7899.89	0.59	0.57
`HistVol`^{r}	Naïve^{r}	8.82	3.97	1.38	$4.0e+00$	1.56	7899.89	0.69	0.86

The "model family" is what I often refer to as the "model category" of each model. "MSE_1"/"$QLIKE_1$" is the model's loss in the period of 2005-2006 relative to the loss of the `HAR-RV` in 2005-2006. "MSE_2"/"$QLIKE_2$" is the model's loss in the period of 2007-2010 relative to the loss of the `HAR-RV` in 2005-2006. "Avg_{MSE_2}"/"Avg_{MSE_2}" is the average relative loss of the MODEL FAMILY. "$\hat{p}_{MSE}$"/"$\hat{p}_{QLIKE}$" are the MCS p-values for the model ("Model name") using the MSE and QLIKE loss functions, respectively. Empty values ("—") in those p-values signify that the model was excluded from the set of superior models (please refer to section 4.4 for more details on the MCS test).