Essays in Computational Statistics with Applications to Volatility Forecasting and Forecast Combination

DISSERTATION
of the University of St. Gallen,
School of Management,
Economics, Law, Social Sciences
and International Affairs
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Simon D. Knaus

from

Hemberg (St. Gallen)

Approved on the application of

Prof. Francesco Audrino, PhD

and

Prof. Marcelo Medeiros, PhD

Dissertation no. 4224

D-Druck Spescha, St. Gallen, 2014

The University of St. Gallen, School of Management, Economics, Law, Social Sciences and International Affairs hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St. Gallen, September 10, 2013

The President:

Prof. Dr. Thomas Bieger

Acknowledgements

I am deeply indebted to Prof. Dr. Francesco Audrino for taking me aboard as his PhD student. To pursue a PhD under his supervision has been a great opportunity. He sparked my interest in many topics in the vast field of computational statistics and their applications to problems in financial econometrics. His openness allowed me to venture into own research projects, yet, he has always been available to discuss and critically comment my work in this field. Without his valuable help my PhD would not have been possible.

I would also like to express my heartfelt gratitude towards the senior faculty of the M+S: Prof. Dr. Enrico De Giorgi, Prof. Dr. Matthias Fengler, Prof. Dr. Daniel Buncic, and Prof. Dr. Lorenzo Camponovo have always answered my knocking on their doors and provided additional opinions to my personal view.

Needless to mention my current and former fellow PhD students: Tatiana, Kameliya, Andreas, Dominik, Constantin, Hans, Pirmin, Daniel, Katja, Markus, Björn, Dewi, Fabian, Anselm, Yujia, Jacopo (in true random order), I thank you all very much for the discussions regarding research and beyond.

A big "thank you" goes also to all my friends outside the ivory tower: I have much appreciated the time on tatamis and mountain tops. A quintessential ingredient to my PhD.

Before embarking on my PhD a friend of mine, who had just completed his PhD, congratulated me on my choice to pursue this avenue. He also mentioned that it is important to have a good personal surrounding. It is only at the end of my PhD that I understand the gravity of his advice. I owe a lot to my family. The ongoing and unconditional support which has ultimately led to my PhD is unparalleled.

Last, but by far not least, Sandrina, who provided support, assistance, guidance, and much-needed energy to complete this journey, thank you!

May 2013 Simon D. Knaus

Contents

Summary **viii**

Zusammenfassung **ix**

List of Figures **ix**

List of Tables **xii**

Preface **1**

1 Lassoing the HAR Model **5**

1.1 Introduction 7

1.2 Theoretical Foundation 10

1.3 Empirical Application 15

1.4 Conclusions and Further Research 31

2 Markets from East to West, News and Volatility Spillovers **35**

2.1 Introduction 37

2.2 Volatility Spillovers 40

2.3 News Arrival 51

2.4 Augmented Linear Models 58

2.5 Discussion and Conclusion 73

3 Micro-level Expert Forecasts **75**

3.1 Introduction 77

3.2 Methods 81

3.3 Data and Procedures 86

3.4 Results 91

3.5 Conclusion 98

Bibliography **99**

Appendices **109**

A Lassoing the HAR model **111**

A.1 Proof of Theorem 2 111

A.2 Log-Transformed Volatilities 119

A.3 Robustness 121

A.4 Risk Management Application 126

A.5 Mincer Zarnowitz Regressions 128

A.6 Implementation 131

B Markets From East to West **133**

B.1 Data Description 133

B.2 Robustness 140

B.3 Implementation 146

C Micro-level Expert Forecasts **149**

C.1 Additional Results 149

C.2 Detailed Methods 151

C.3 Implementation 153

Summary

The first paper investigates the dynamics of realized volatility. Recent notable advances that perform well include the heterogeneous autoregressive (HAR) model which is economically interpretable but still easy to estimate. It also features good out-of-sample performance and has been extremely well received by the research community. We present a data driven approach based on the absolute shrinkage and selection operator which should identify the aforementioned model. We prove that the lasso indeed recovers the HAR model asymptotically if it is the true model, and we present Monte Carlo evidence in finite sample.

The second paper investigates the role of volatility spillovers, macroeconomic news, intraweek seasonality and the leverage effect for realized volatility models. To understand the role of news existing models, augmented by this new combined information set, are revisited. The least absolute shrinkage and selection operator is employed in a vector autoregressive setting to assess the relevance of volatility spillovers to modeling the S&P 500's realized volatility. A combined model is then proposed which features a data-driven selection of regressors to include in a model of realized volatility. These models are compared in a strict out-of-sample prediction comparison together with a value-at-risk application. A superior performance of models augmented with this new information set is witnessed in the prediction comparison. Compared to existing models, a considerably shorter lag structure delivers already good forecasting performance.

The third paper investigates forecast combination in the field of macroeconomics. While macroeconomic survey forecasts are widely available at the level of individual experts, it is not clear how to optimally combine a set of forecasts to a "consensus" prediction. This is mainly due to the characteristics of the data, such as the large-dimensional predictor space, many missing values, and potential individual and aggregate level biases of the survey forecasts. We argue that regression trees are very well adapted to these features and propose to use them as a novel forecast combination device. Our empirical analysis of data from the Philadelphia Fed's Survey of Professional Forecasters demonstrates that in combination with bagging, tree-based forecast combination outperforms equally weighted combination for the majority of time series and forecast horizons.

Zusammenfassung

Der erste Aufsatz der vorliegenden Dissertation untersucht die Dynamik der realisierten Volatilität. Ein erfolgreiches Modell in diesem Bereich ist das sogenannte heterogene autoregressive Modell (HAR), das konzeptionell einfach und gut für Vorhersagen geeignet ist. Eine neue Herangehensweise basierend auf dem *Operator der kleinsten absoluten Schrumpfung und Auswahl* ermöglicht es, das HAR Modell von einem Modellwahl Standpunkt her zu betrachten. Es wird gezeigt, dass die Modellwahl asymptotisch das wahre Modell identifizieren könnte. Zusätzlich werden simulierte Resultate im nicht-asymptotischen Bereich präsentiert, welche die Modellwahlgüte unterstreichen. Zusammenfassend kann gesagt werden, dass das HAR Model wohl nicht als wahres Modell identifiziert wird, das gewählte Model aber nicht vom HAR Modell zu unterscheiden ist, falls Vorhersagegüte als Vergleichskriterium herangezogen wird.

Der zweite Aufsatz untersucht den Einfluss von externen Faktoren auf die Dynamik von Modellen für die realisierte Volatilität. Diese externen Faktoren umfassen Volatilitätsübertragung zwischen Märkten, die Bekanntgabe von makroökonomischen Kenngrössen, den Hebeleffekt sowie innerwöchentliche Saisonalitäten. Um die Rolle der einzelnen Faktoren besser zu verstehen wird wiederum der Operator der kleinsten absoluten Schrumpfung und Auswahl herangezogen. Zusätzlich werden diese Resultate mit existierenden Modellen verglichen, um die Relevanz der genannten Faktoren auf die Modellierung der realisierten Volatilität des S&P 500 Index abzuschätzen. Es kann festgehalten werden, dass ein um diese Informationen erweitertes Modell tatsächlich bessere Vorhersagen für die realisierte Volatilität liefert.

Der dritte Aufsatz untersucht die Kombination von Expertenprognosen für makroökonomische Daten. Obwohl individuelle Expertenprognosen sehr verbreitet sind ist es a priori nicht klar, wie diese zu einer aggregierten Vorhersage kombiniert werden können. Dies ist hauptsächlich speziellen Eigenschaften der Daten, wie z.B. hoch-dimensionale Prädiktorräume und fehlende Werte, geschuldet. Wir kommen zum Schluss, dass Regressionsbäume wohlgeeignet sind um diese Eigenschaften der Daten auszunutzen und eine aggregierte Prognose zu erhalten. Die empirische Untersuchung fusst auf Daten der Zentralbank von Philadelphia und deren Umfrage unter professionellen Prognostikern. In dieser zeigt sich, dass Regressionsbäume sowie deren robustifizierte Variante in der Tat in der Lage sind, aggregierte Prognosen zu liefern, die in vielen Fällen dem einfachen Mittelwert überlegen sind.

List of Figures

1.1 Autocorrelation function for $\log \mathrm{RV}_t$ series. 16

1.2 HAR versus lasso coefficients with all predictors 22

1.3 Stability of selected regressors for HDI 24

1.4 Stability of Lasso selected regressors for all assets 25

1.5 Actual hit ratios. 30

2.1 Index trading hours 42

2.2 Quantile-quantile plot of (standardized) scaled logarithmic returns 67

3.1 Unemployment forecasts 80

3.2 Regression tree 83

3.3 MSPE decomposition 97

A.1 Coefficient of determination along power transform 120

A.2 Autocorrelation function for $\log \mathrm{RV}_t$ series using MedRV. 122

A.3 HAR versus lasso coefficients with all predictors using MedRV estimator . . 123

A.4 Stability of Lasso selected Regressors for all assets using MedRV estimator . 124

A.5 Kernel density estimates of standardized log-returns for pre-crisis (PC) and full sample (FS) against normal distribution. 128

A.6 Mincer Zarnowitz Estimates (Intercept) 129

A.7 Mincer Zarnowitz Estimates (Slope) 130

A.8 Joint hypothesis p-values 131

B.1 Distribution of unconditional $\log \mathrm{RV}_t$ 137

B.2 Autocorrelation over time 138

B.3 Full day versus partial day realized volatility 138

B.4 Distribution of standardized returns 145

B.5 Distribution of standardized (forecasted) returns 147

List of Tables

1.1 Descriptive statistics of $\log \mathrm{RV}_t$ series 17

1.2 HAR estimates versus lasso estimates 20

1.4 Out-of-sample comparison 27

1.5 Diebold-Mariano (Diebold & Mariano 1995) tests of equal predictive ability 28

1.3 Percentage of HAR coefficients recovered 33

2.1 Structural and reduced Form Estimates 49

2.2 In-sample estimates of model classes 1 to 3 56

2.3 Out-of-sample results 62

2.4 Out-of-sample value-at-risk forecasts 71

3.1 Data description 88

3.2 Notation 90

3.3 Relevant information sets 91

3.4 Full evaluation sample 94

A.1 Diebold-Mariano tests of equal predictive ability 125

A.2 VaR violations (hit ratios) 127

B.1 Descriptive statistics 137

B.2 Regression estimates full versus partial day 139

B.3 Out-of-sample forecasts with varying training window widths 141

B.4 Out-of-sample results using the naive estimator 142

B.5 Out-of-sample forecasts with varying training window widths using naive estimator 144

B.6 Out-of-sample value-at-risk forecasts using naive estimator 146

C.1 Bagged predictors (full sample) 150

C.2 Bagged predictors (pre-crisis) 150

C.3 Pre-crisis evaluation sample 154

Preface

“Essays in Computational Statistics with Applications to Volatility Forecasting and Forecast Combination” presents research at the intersection of computational statistics and (financial) econometrics. The field of computational statistics has seen an ever increasing popularity in recent decades. The advent of data and affordable computing power helped to spread techniques developed in this field. Most methods originating from computational statistics do indeed require substantial computational efforts when compared to more traditional approaches. In recent years, many of these techniques have also found applications in (financial) econometrics and have enlarged the set of tools to approach relevant research questions. Many methods originating from the field of computational statistics are indeed able to better capture properties of a certain problem; the simplicity and transparency of traditional approaches however, some of them dating back to Legendre and Gauß (Legendre 1805, Gauß 1887), cannot always be maintained. While some scholars summarize these methods as machine learning or statistical learning, others doubt that anything can be *learned* by simply crunching more data with more computing power.

An often voiced critique in this regard goes by the name of “data snooping” or “data mining”. To answer this critique all three essays strictly maintain the fundamental idea of separating the in-sample data (exploring the data) from the out-of-sample data (testing the conclusion drawn in-sample on previously unseen data). With this approach, I consider the application of techniques from the field of computational statistics to problems in (financial) econometrics truly helpful to gain a deeper understanding of a problem. An often heard quote, which is attributed to William E. Deming (Hastie, Tibshirani & Friedman 2009), is "In God we trust, all others bring data". The uncovering of this innate message of observed data complemented with antecedent research, established concepts in the respective field, and due care in the interpretation of the results makes these techniques valuable means to gain new insights. Consequently, it is possible to *learn* something about the problem and thus potentially push the research frontier by a small amount.

The three essays constituting this dissertation all share that they feature an application of a particular approach in computational statistics to (financial) econometrics. The use of these tools is guided by the ideas outlined above.

The first two chapters deal with the idea of volatility. As opposed to financial returns, which, under common assumptions, do not allow for a consistent estimation of the expected return in finite time, volatility can be estimated consistently if there are enough observations. This property underlies the analysis of *realized volatility*. Realized volatility, drawing on the plethora of data from financial markets, exploits this feature to construct a measure for volatility based on intraday data. With a series of realized volatility measures at hand, the problem of modeling its dynamics can be approached.

The first chapter, coauthored with Francesco Audrino, exploits and extends recent advancements in model selection to the realm of realized volatility modeling. While model selection is computationally costly with traditional means, the least absolute and selection operator (lasso), a recent contribution in the field of computational statistics, which is increasingly used in econometrics as well, considerably lowers the computational burden and provides model selection in this setting. With this tool at hand we can approach the problem of verifying whether a popular model, the heterogeneous autoregressive model (HAR) for realized volatility, can actually be the true model. This chapter concludes that this model may not be the true model. Nonetheless it captures a linear footprint of the true volatility dynamics, as does the lasso, which also allows modelling realized volatility series.

The second chapter extends ideas developed in the first chapter along different lines. Again, realized volatility is studied using the lasso. However, the focus is different: In contrast to the first chapter, where the ultimate goal is model selection, the paper is led by the goal of volatility forecasting. Drawing on existing research, which looks into factors affecting the volatility series such as spillovers or the arrival of news, I employ the lasso in a multivariate as well as univariate setting to devise a model which extends again the HAR model. This newly proposed model, which features volatility observations from geographically preceding markets, the arrival of macroeconomic news, weekly seasonality, and a leverage effect, does fare well in an out-of-sample comparison. Hence, I conclude that indeed augmenting the HAR model with these factors is beneficial to realized volatility forecasting.

The last chapter deviates from the path outset by the first two chapters. However, the particularities of the data and the problem investigated in Chapter 3 requires again methods from computational statistics. As opposed to the first two chapters, where data is available in abundance, I investigate together with Fabian Krüger, the benefit of employing non-parametric methods in macroeconomic forecast combination. Recently, Alex Tabarrok, a scholar in macroenomics, let the world know "Ultimately, the problem with macro economics is simple: small data." via social media (Tabarrok 2013). Although, this statement is of much broader nature, we also face this problem in our study. Combining quarterly forecasts in a non-parametric way is usually impeded by, amongst other problems, the curse of dimensionality. To address the specificities of this data, we employ regression trees as combination device and study the effects of individual level data and bias correction. We contrast our results with well-established combination schemes and conclude in summary that robustified regression trees may help to improve the forecasts of certain macroeconomic data by mitigating the bias inherent to expert predictions whereof other combination schemes suffer.

Chapter 1

Lassoing the HAR Model: A Model Selection Perspective on Realized Volatility Dynamics

Francesco Audrino
Simon D. Knaus

Abstract

Realized volatility computed from high-frequency data is an important measure for many applications in finance. However, its dynamics are not well understood to date. Recent notable advances that perform well include the heterogeneous autoregressive (HAR) model which is economically interpretable and but still easy to estimate. It also features good out-of-sample performance and has been extremely well received by the research community.

We present a data driven approach based on the absolute shrinkage and selection operator (lasso) which should identify the aforementioned model. We prove that the lasso indeed recovers the HAR model asymptotically if it is the true model, and we present Monte Carlo evidence in finite sample. The HAR model is not recovered by the lasso on real data. This, together with an empirical out-of-sample analysis that shows equal performance of the HAR model and the lasso approach, leads to the conclusion that the HAR model may not be the true model but it captures a linear footprint of the volatility dynamics.

JEL: C58, C63, C49

Keywords: Realized Volatility, Heterogeneous Autoregressive Model, Lasso, Model Selection

1.1 Introduction

Volatility of financial assets is of great importance to many applications in finance. Reliable estimates and forecasts are key for risk management and asset allocation. As opposed to returns series, financial volatility is predictable and has received great attention in the financial econometrics research community. The seminal paper of Bollerslev (1986) introducing the generalized autoregressive conditional heteroscedasticity (GARCH) model for conditional volatility has thus sparked an even greater interest in volatility modeling. The GARCH model has become extremely popular and despite various extensions and modifications the basic GARCH(1,1) fares well as a prediction device for conditional volatility in an out-of-sample forecast comparison (Hansen & Lunde 2005). While Bollerslev's (1986) GARCH model is able to capture stylized facts of volatility series (e.g., volatility clustering), its estimation still relies on daily observations and thus potentially discards intraday information. The advent of high-frequency data (with frequencies as high as tick-by-tick) has ignited a new line of research pioneered by Andersen, Bollerslev, Diebold & Labys (2001) and Barndorff-Nielsen & Shephard (2002*b*) among others. The results of their work has rendered the thus far unobservable daily volatility observable by means of asymptotic arguments:

Suppose that an asset's log price obeys the dynamics $\mathrm{d}X_t = \mu_t \mathrm{d}t + \sigma_t \mathrm{d}W_t$ where W_t is a Brownian motion, σ_t the instantaneous volatility and μ_t the instantaneous drift term. One can then show that $\operatorname{plim}_{\delta\to 0} \sum_{t_i}(X_{t_{i+1}} - X_{t_i})^2 = \int_0^T \sigma_t^2 \mathrm{d}t$ where $\delta = \sup\{t_{i+1} - t_i\}$, i.e., the sum of squared returns converges to the integrated variance (over a day) as the sampling frequency increases.[1] An estimator of $\int_0^T \sigma_s^2 \mathrm{d}s$ is thus given by $\sum_{i=1}^N (X_{t_{i+1}} - X_{t_i})^2$ where $t_1, \ldots, t_N$ is an appropriate sampling frequency and is denoted RV_t, where t refers to the day. RV_t is called *realized variance*, and its squareroot $\sqrt{\mathrm{RV}_t}$ is referred to as *realized volatility*. An overview of variants of the aforementioned estimator and their corresponding assumptions is collected in McAleer & Medeiros's (2008) review on realized volatility.

[1] It is known that this naive estimator of $\int_0^T \sigma_t^2 \mathrm{d}t$ is biased under e.g., microstructure noise (the observable return process $Y_{t_i} = X_{t_i} + \varepsilon_{t_i}$ is contaminated with noise) or if the log price process is a jump-diffusion ($\mathrm{d}X_t = \mu_t \mathrm{d}t + \sigma_t \mathrm{d}W_t + \mathrm{d}J_t$ where J_t is a finite activity jump process).

Since the goal of this work is to investigate the dynamics of the realized variance and *not* the estimation itself we can thus – with daily realized variance at hand – approach the problem of *modeling* realized variance.

It has been observed that the time series $\{RV_t\}_{1\leq t\leq T}$ exhibits some distinct features such as a near log-normal unconditional distribution as well as a slowly decaying autocorrelation function which is often termed "long memory": These findings appear to be robust across different asset classes and evidence has been reported for exchange rates (Andersen, Bollerslev, Diebold & Labys 2001), index futures (Areal & Taylor 2002, Thomakos & Wang 2003), as well as for individual stocks (Andersen, Bollerslev, Diebold & Ebens 2001).

To address these characteristics of the realized variance time series, different approaches have been put forward, most prominently fractionally integrated ARMA models (ARFIMA) and the heterogeneous autoregressive (HAR) model for realized volatility introduced by Corsi (2009). The HAR model not only allows for an economic interpretation of the proposed dynamics, but also allows for an easy estimation and is thus highly appreciated and widely used within the research community.

The contribution of this paper is to shed more light on the underlying dynamics as advocated by Corsi's (2009) HAR model which in essence claims tomorrow's realized variance to be a sum of daily, weekly, and monthly averages of realized variances that can each be attributed to specific investment behaviors. The question we are aiming to answer relates to how much these frequencies (daily, weekly, monthly) are really inherent to the data and if we can identify them from a model selection perspective.

Model selection plays a crucial role in determining a model for forecasting. Oftentimes model selection can be extremely costly from a computational perspective and may already become infeasible within the class of linear models (an exhaustive search over p lags already requires 2^p comparisons and thus grows exponentially). An important contribution in terms of model selection within the class of linear models was made in Tibshirani (1996) where the Least Absolute Shrinkage and Selection Operator (lasso) was introduced. The lasso, a shrunk regression, performs shrinkage and selection at a time and is yet computationally affordable. Although originally the lasso was mostly noticed by the computational statistics community, researchers in econometrics are increasingly using it. Most recently, conditions under which the lasso gives consistent results have also been established in time series econometrics (Nardi & Rinaldo 2011), and applications of the lasso are also found in Park & Sakaori (2013).

Despite the great popularity and appreciation of the HAR model there has been little work investigating the validity of the structure as proposed by the HAR model. Although most work is done in the direction of extending the HAR model (see the recent review of Corsi, Audrino & Renò (2012)) there is a notable exception: Craioveanu & Hillebrand (2010) investigate the structure of the HAR model and find no benefit in allowing for a more flexible structure of lag selection. However, their result is based on an exhaustive search over HAR-like models but varying aggregation frequencies.

It is along these lines that this paper adds to the literature. We present a methodologically sound way of recovering the HAR model. We show that under the assumption that HAR model is the true model, we can apply the lasso and should recover the structure as implied by the HAR model. To this end we investigate how far Nardi & Rinaldo's (2011) result can be extended for the special case of the HAR model. Moreover, we investigate if the lasso can be used for forecasting realized variance from a purely statistical point of view as well as measuring outperformance from a more economically relevant point of view via a risk management application. We find no substantial superiority of either the HAR model or the lasso when it comes to out-of-sample forecasting.

In summary, we have reason to believe that the HAR model might not be the true model. However, it captures a linear footprint of the true underlying variance dynamics which appear to change over time, thus casting some doubt on the appropriateness of the HAR as a global model for realized variance.

The rest of the paper is structured as follows: Section 1 introduces the HAR model in more detail, relates it to the autoregressive class of time series models and shows how the lasso can be used in this context. Section 2 features an empirical application of the proposed model selection approach, a Monte Carlo study, as well as an out-of-sample comparison of the HAR versus the lasso. Section 3 discusses the results and further research and then concludes.

1.2 Theoretical Foundation

1.2.1 The HAR Model

The HAR model as introduced in Corsi (2009) enjoys great popularity: It allows for an economic interpretation, has good forecasting performance, and is still easy to estimate. There are numerous variants and modifications of the HAR model (Corsi et al. 2012), however we restrict our attention to the original model to keep a clear focus on the actual volatility dynamics. We thus intentionally ignore other transient effects (such as the leverage effect) that may be embedded in a HAR framework as well.

Let for this purpose $\mathrm{RV}_t^{(d)}$ be an estimate of daily realized variance. Then, the HAR model postulates that

$$\log \mathrm{RV}_{t+1}^{(d)} = c + \beta^{(d)} \log \mathrm{RV}_t^{(d)} + \beta^{(w)} \log \mathrm{RV}_t^{(w)} + \beta^{(m)} \log \mathrm{RV}_t^{(m)} + \omega_{t+1}, \tag{1.1}$$

where (with a slight abuse of notation) $\log \mathrm{RV}_t^{(w)} = \frac{1}{5}\sum_{i=1}^{5} \log \mathrm{RV}_{t-i+1}^{(d)}$ and $\log \mathrm{RV}_t^{(m)} = \frac{1}{22}\sum_{i=1}^{22} \log \mathrm{RV}_{t-i+1}^{(d)}$ are the weekly and monthly averages of daily log realized variances, and ω_{t+1} is an innovation. Once these average log-variances are known, the model can be consistently estimated by traditional least squares to obtain estimates for c, $\beta^{(d)}$, $\beta^{(w)}$, and $\beta^{(m)}$.

In other words, the conditional expectation of tomorrow's log-realized variance is the weighted sum (plus an intercept) of daily, weekly, and monthly log-realized volatilities.[2] For the remainder of the paper we assume the HAR model to be causal as well as $\beta^{(d)}$, $\beta^{(w)}$, $\beta^{(m)}$ to be positive. These assumptions are by no means restrictive: First they comply

[2] We comment further on the use of log-realized volatilities in Section 1.3.1.

with the view put forward in the original work as outlined below, second, if estimating the HAR on empirical data, the coefficients are always found to be positive.

The different aggregation frequency can then be seen as a heterogeneous agent model where heterogeneity is induced by the different time horizons and can be casted into an information cascade view. Hence, the weighted average perspective appears reasonable and positiveness of the coefficients follows.

Clearly, the HAR model is simply a constrained AR(22) model, as it has already been noted by Corsi (2009), i.e., we can write

$$\log \mathrm{RV}_{t+1}^{(d)} = \phi^{\mathrm{HAR}} + \sum_{i=1}^{22} \phi_i^{\mathrm{HAR}} \log \mathrm{RV}_{t-i+1}^{(d)} + \omega_{t+1} \tag{1.2}$$

where the restrictions as imposed by (1.1) require

$$\phi_i^{\mathrm{HAR}} = \begin{cases} \beta^{(d)} + \frac{1}{5}\beta^{(w)} + \frac{1}{22}\beta^{(m)} & \text{for } i = 1 \\ \frac{1}{5}\beta^{(w)} + \frac{1}{22}\beta^{(m)} & \text{for } i = 2, \dots, 5 \\ \frac{1}{22}\beta^{(m)} & \text{for } i = 6, \dots, 22. \end{cases} \tag{1.3}$$

A direct specification test is obviously testing the restrictions as collected by (1.3). Given the high number of restrictions a rejection of these is not surprising. However, in the original work Corsi argues that this can well be attributed to specific properties of the time series. However, there is already some preliminary indication that indeed the HAR model may fail to fully capture the effects present in the data.

1.2.2 The lasso as model selection device

The lasso was introduced in Tibshirani (1996) and is frequently used in the field of computational statistics and machine learning. In recent years, the lasso in general as well as the lasso as model selection device has also been found in Econometrics (Kock 2012, Leeb & Pötscher 2005). The lasso is computationally very efficient and renders model selection with a high number of predictors feasible. As opposed to the 2^p comparisons that are required in an exhaustive search over p predictors, the lasso employs a highly efficient algorithm which provides estimates and model selection jointly (Friedman, Hastie & Tibshirani 2010) at affordable computational costs.

The lasso as originally introduced by Tibshirani covered the cross-sectional case: Let $x_i = (x_{i1}, \ldots, x_{ip})'$ be predictor variables and y_i responses. Under the assumption that the predictors are standardized, the lasso estimator of the model

$$y_i = \alpha + \phi' \cdot x_i + \epsilon_i \tag{1.4}$$

is obtained as

$$(\hat{\alpha}^{\text{lasso}}, \hat{\phi}^{\text{lasso}}) = \arg\min_{\alpha,\phi} \left\{ \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} \phi_j x_{ij} \right)^2 \right\}$$
$$\text{subject to} \sum_{j=1}^{p} |\phi_j| \leq t \tag{1.5}$$

where t is a tuning parameter. Since $\hat{\alpha}$ is independent of t it will always be equal to $\bar{y}$ and it is thus generally assumed that $\bar{y} = 0$ and α is dropped from the minimization. It can be seen (Tibshirani 1996) that (1.5) is equivalent to the Lagrangian form given as

$$\hat{\phi}^{\text{lasso}} = \arg\min_{\phi} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \phi_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\phi_j| \right\} \tag{1.6}$$

with a one-to-one correspondence between λ in (1.6) and t in (1.5). The powerful feature of the lasso is now induced by the L^1-norm of the penalty. The lasso solution will be sparse, since some ϕ_js will be set exactly to zero (as opposed to for instance ridge regularization in Hastie et al. (2009) where sparsity of the solution is lost due to the L^2-geometry of ridge).

A question of utmost importance is how reliable is the lasso in the sense that it sets the true zero coefficients to zero. Typically, this is what is captured by *model selection consistency*. The following definition adopts the view of Nardi & Rinaldo (2011). For an overview and weaker form of this, the reader is referred to Bühlmann & Van De Geer (2011).

Definition 1. *Let* $y_i = \phi' \cdot x_i + \epsilon_i$ *with* $\phi^0 = [\phi_1^0, \ldots, \phi_p^0]'$, $\text{sgn} : \mathbb{R} \to \{-1, 0, 1\}$ *and define* $\text{sgn}(\phi) = (\text{sgn}(\phi_1), \ldots, \text{sgn}(\phi_p))'$*. Then an estimator* $\hat{\phi}_n$ *is said to be* model selection consistent *if*

$$P(\text{sgn}(\hat{\phi}_n) = \text{sgn}(\phi^0)) \to 1 \text{ for } n \to \infty. \tag{1.7}$$

The above model selection consistency definition meets our requirement that if there is an estimator producing $\hat{\phi}_n$ which is *model selection consistent* it will eventually only retain the true non-zero coefficients supp ϕ^0.

An extension of the lasso as well as proof for which conditions the lasso is model selection consistent is given in Zou (2006). Zou introduces the adaptive lasso which allows for a more flexible penalization, i.e.,

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \phi_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \lambda_j |\phi_j| \right\} \tag{1.8}$$

where λ_j are adaptive weights. It can be shown (Zou 2006, Bühlmann & Van De Geer 2011) that in fact the adaptive lasso relaxes the assumptions for the model selection consistency of the lasso.

An important extension of this strand of literature has been made by Nardi & Rinaldo (2011): Nardi & Rinaldo show that properties already well-established in the cross sectional case carry over to the time series case of an AR(p) process.[3] More precisely, they establish that under some assumptions, a version of the adaptive lasso is model selection consistent. Suppose that X_t is a causal Gaussian AR(p) process, i.e.,

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \epsilon_t$$

where ϵ_t is i.i. $\mathcal{N}(0, \sigma^2)$-distributed. Define $S = \{j, \phi_j \neq 0\} \subset \{1, \ldots, p\}$ the active set, $S^c = \{1, \ldots, p\} \setminus S$ the non-active set, and $\Gamma_{XY} = \text{Cov}(X, Y)$ the covariance matrix of a vector X and Y. Consequently, Γ_{SS} is the square covariance matrix of the active predictors and

[3] Note that Definition 1 is by no means limited to the cross-sectional case and translates directly to the time series regression variant.

Γ_{S^cS} is the covariance matrix of the predictors in the non-active set (given as $\{X_{t-j},\ j \in S^c\}$) with the predictors in the active set (given as $\{X_{t-j},\ j \in S\}$). They then proceed and prove the following theorem (Nardi & Rinaldo 2011, Theorem 3.1):

Theorem 1. *Consider the AR(p) settings described above. Assume that*

(i) *there exists a finite positive constant* $C_{\max}$ *such that* $\|\Gamma_{SS}^{-1}\|_\infty \leq C_{\max}$*;*

(ii) *there exists a* $\delta \in (0,1]$ *such that* $\|\Gamma_{S^cS}\Gamma_{SS}^{-1}\|_\infty \leq 1-\delta$.

Further assume that the asymptotic properties for λ_n *and* $\lambda_{n,j}$ *as given in Nardi & Rinaldo (2011, Theorem 3.1) hold.*

Then, the lasso estimator is model selection consistent in the sense of Definition 1.

Condition (ii) of the above theorem is found throughout the model consistency literature for the lasso. Typically this condition is called the *irrepresentable condition* as introduced in Zhao & Yu (2006). Nardi & Rinaldo show that a causal Gaussian process satisfies the assumptions of Theorem 1 and the lasso is thus *model selection consistent* for this class of models.

1.2.3 Lassoing the HAR model

Theorem 1 states that the lasso is indeed model selection consistent for causal AR(p) processes with Gaussian innovations. If we assume that ϵ_t in (1.1) is Gaussian we can readily use the lasso to try to recover the HAR model embedded in an AR(p) process with $p > 22$. The lasso should then detect[4] $S = \{1, 2, \ldots, 22\}$ and $S^c = \{23, 24, \ldots, p\}$ since any other lagged value should be irrelevant if the HAR model is the true data generating process (DGP).

[4] In the sense of setting the non-active coefficients to zero.

The assumption of Gaussianity of the error may appear strong at first sight. However, the HAR model is usually estimated using quasi-likelihood which in turn also assumes Gaussianity. An even stronger argument is given below and proved in the appendix. Under the assumption that the HAR model is the true DGP, we precisely know the dynamics and can prove (ii) of Theorem 1 directly without relying on Gaussianity. This can then be used in Zhao & Yu's (2006) result which relaxes the assumptions of Gaussianity of the innovations. The relaxation on the distribution of the error term comes at the price of keeping S and S^c fixed; the lasso literature generally differentiates between a $p = |S|$ growing with n or p fixed. Theorem 1 above addresses the case where p is allowed to grow, our contribution below however requires p to be fix:

Theorem 2. *Under the assumptions that the DGP is as given in* (1.1) *is causal and the innovation has a finite fourth moment, S^c is held fixed, then lasso is model selection consistent in the sense of Definition 1.*

The complete argument and proof is given in Appendix A.1.

1.3 Empirical Application

In this section we illustrate our approach of identifying the HAR model via the lasso using nine assets traded on the New York Stock Exchange. For each of these stocks we compute a realized variance measure using Zhang, Mykland & Aït-Sahalia's (2005) two-time scales estimator (using a frequency of 10 minutes) to obtain a series of daily realized variance measures.[5] These measures are then used to estimate the HAR model in-sample and contrast it with estimates as obtained by the lasso procedure described in Section 1.2. We also compare the lasso's forecasting performance to the performance of the HAR out-of-sample. To rule out any doubt that these findings are dependent on a specific realized variance estimator we also report a summary of results using Andersen, Dobrev & Schaumburg's (2010) MedRV estimator in Appendix A.3. The key descriptive properties of the data are summarized in Fig. 1.1 and Tab. 1.1.

[5] We adhere to the suggestion put forward in Corsi (2009) and use annualized returns in percentage points.

Note that we obviously only forecast one day ahead realized variance since our argument is based on the original specification of the HAR model. One could of course address the question whether the lasso is also well suited to forecast realized variance at longer horizons (weekly, monthly); this however would be a purely empirical exercise and is beyond the scope of this paper.

1.3.1 Data Description

We use intraday data of Alcoa, Inc. (AA), Citigroup, Inc. (C), Hasbro Inc. (HAS), The Home Depot, Inc. (HDI), Intel Corporation (INTC), Microsoft Corporation (MSFT), Nike Inc. (NKE), Pfizer Inc. (PFE), and Exxon Mobil Corporation (XOM) from Jan 2, 2001 to Nov 15, 2010. These intraday data are then used to compute an estimator of daily realized variance using Zhang et al.'s (2005) two-time scales estimator. In total we have 2483 observations of realized variance measures.

Figure 1.1: Autocorrelation function for $\log \mathrm{RV}_t$ series.

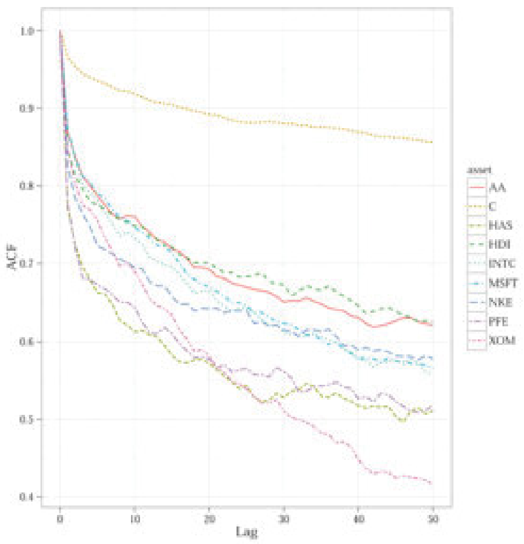


Figure 1.1 (a)

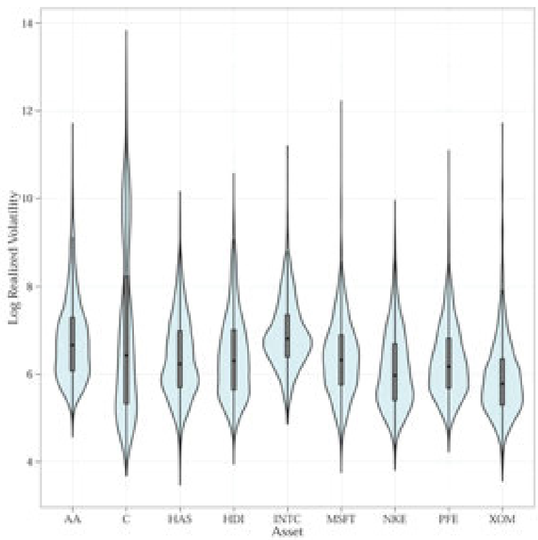


Figure 1.1 (b)

Panel (a) shows the autocorrelation function for the 9 $\log \mathrm{RV}_t$ series. Panel (b) shows a violin plot (Hintze & Nelson 1998) of the unconditional $\log \mathrm{RV}_t$

Table 1.1: Descriptive statistics of $\log \mathrm{RV}_t$ series

	AA	C	HAS	HDI	INTC	MSFT	NKE	PFE	XOM
Mean	6.80	6.96	6.36	6.42	6.90	6.37	6.10	6.30	5.90
SD	0.95	1.98	0.93	0.98	0.78	0.89	0.90	0.85	0.89
Kurtosis	4.06	2.61	3.06	3.32	3.82	4.40	3.06	4.18	6.15
Skewness	0.92	0.78	0.46	0.69	0.58	0.58	0.58	0.83	1.19
Median	6.67	6.43	6.23	6.30	6.81	6.33	5.97	6.18	5.78
25%-quantile	6.07	5.33	5.70	5.65	6.40	5.76	5.41	5.68	5.30
75%-quantile	7.30	8.21	6.99	7.01	7.35	6.90	6.70	6.82	6.35

Although using the log to transform the realized variance is standard in the literature, we briefly comment explicitly on this in Appendix A.2 for the HAR model. In what follows we always assume the use of log realized variance when speaking of realized variance unless otherwise stated.

Consistent with the existing literature we witness slowly decaying autocorrelation functions in Fig. 1.1 (a) for all assets. This is most pronounced for Citigroup, Inc. The same stock also exhibits particularities in the unconditional distribution of $\log \mathrm{RV}_t$ as can be seen from Tab. 1.1 and Fig. 1.1 (b): While all other stocks show excess kurtosis, Citigroup Inc. only has a kurtosis of 2.61. We suspect the market turmoil of the financial crisis to be the root of this abnormal picture. Following this train of thought, we also report the actual returns in Fig. A.5 in the appendix where an extremely high excess kurtosis for the log returns of Citigroup Inc. can be observed.

1.3.2 In-sample Evaluation

To address the question whether the HAR model is identified by the lasso procedure we define $S^c = \{x_{t-23}, \ldots, x_{t-100}\}$.[6] Since λ in (1.6) is a tuning parameter and the results of Theorem 1 only hold asymptotically we proceed as suggested in the literature (Nardi &

[6] The choice of S running up to 100 is arbitrary. However, the results are not sensitive to the choice of the maximal lag, as for instance the results remain almost identical for a maximal lag of 50

Rinaldo 2011, Section 4.1.) and choose $\lambda_j = 1$ for all j and $\lambda = \sqrt{\frac{\log n \log p}{n}}$ and can thus expect $\hat{S}$ as obtained by $\hat{\phi}^{\text{lasso}}$ to be sparse in $\{1, \ldots, 100\}$.

Two important points should be noted here: First, the lasso does not recover all of the coefficients implied to be non-zero by the HAR as can be inferred from Tab. 1.2. Although near lags are recovered for most assets, lags beyond x_{t-6} rarely get selected by the lasso. Note at this point that a comparison of coefficients in magnitude of the lasso estimates to the HAR estimates cannot be made since the lasso, as a penalized estimator, is biased. Second, sometimes lags far beyond x_{t-22} are selected in the active set as can be seen in Fig. 1.2. Clearly, these lags are zero under the assumption that the HAR model is true.

At this stage it is alaredy apparent that the lasso does *not* fully recover the HAR model, i.e. $\hat{S} \neq \{1, \ldots, 22\}$. To provide further evidence supporting this statement, we conduct analyses which attempt to answer the following two questions: 1. How reliable is the lasso as a model selection device in this specific finite sample setting? 2. How stable are these regressors over time? A thorough answer to these questions is provided in the two subsequent paragraphs.

Table 1.2: HAR estimates versus lasso estimates

	AA		C		HAS		HDI		INTC		MSFT		NKE		PFE		XOM	
Lag	HAR	lasso	HAR	lasso	HAR	lasso	HAR	lasso	HAR	lasso	HAR	lasso	HAR	lasso	HAR	lasso	HAR	lasso
γ_1	0.470	0.418	0.573	0.530	0.403	0.358	0.413	0.370	0.535	0.483	0.488	0.444	0.456	0.408	0.417	0.376	0.497	0.455
γ_2	0.084	0.172	0.073	0.169	0.084	0.121	0.086	0.137	0.074	0.118	0.086	0.125	0.073	0.116	0.073	0.112	0.094	0.130
γ_3	0.084	0.039	0.073	0.001	0.084	0.072	0.086	0.077	0.074	0.003	0.086	0.042	0.073	0.076	0.073	—	0.094	0.033
γ_4	0.084	0.061	0.073	0.065	0.084	0.040	0.086	0.019	0.074	0.062	0.086	0.087	0.073	0.049	0.073	0.067	0.094	0.096
γ_5	0.084	0.029	0.073	0.044	0.084	0.022	0.086	0.050	0.074	0.049	0.086	0.014	0.073	—	0.073	0.039	0.094	0.052
γ_6	0.010	0.010	0.008	0.024	0.012	0.053	0.012	0.052	0.008	—	0.008	0.051	0.013	0.027	0.014	0.040	0.005	—
γ_7	0.010	—	0.008	—	0.012	0.032	0.012	0.012	0.008	—	0.008	—	0.013	0.031	0.014	0.025	0.005	—
γ_8	0.010	—	0.008	—	0.012	—	0.012	0.013	0.008	—	0.008	—	0.013	0.005	0.014	0.005	0.005	—
γ_9	0.010	0.043	0.008	0.056	0.012	—	0.012	0.029	0.008	0.042	0.008	0.040	0.013	0.023	0.014	0.036	0.005	0.039
γ_{10}	0.010	0.056	0.008	0.001	0.012	—	0.012	0.013	0.008	—	0.008	0.002	0.013	0.004	0.014	0.027	0.005	—
γ_{11}	0.010	—	0.008	—	0.012	0.021	0.012	0.030	0.008	—	0.008	0.002	0.013	0.025	0.014	—	0.005	—
γ_{12}	0.010	—	0.008	—	0.012	—	0.012	—	0.008	—	0.008	—	0.013	—	0.014	—	0.005	—
γ_{13}	0.010	—	0.008	—	0.012	0.011	0.012	0.003	0.008	—	0.008	—	0.013	—	0.014	—	0.005	—
γ_{14}	0.010	0.005	0.008	0.005	0.012	—	0.012	—	0.008	—	0.008	0.006	0.013	0.008	0.014	0.030	0.005	—
γ_{15}	0.010	—	0.008	—	0.012	—	0.012	—	0.008	—	0.008	—	0.013	—	0.014	—	0.005	—
γ_{16}	0.010	—	0.008	—	0.012	—	0.012	—	0.008	—	0.008	—	0.013	—	0.014	—	0.005	—
γ_{17}	0.010	—	0.008	—	0.012	—	0.012	—	0.008	—	0.008	—	0.013	—	0.014	—	0.005	—
γ_{18}	0.010	—	0.008	—	0.012	—	0.012	—	0.008	—	0.008	—	0.013	—	0.014	—	0.005	—
γ_{19}	0.010	—	0.008	—	0.012	0.007	0.012	—	0.008	—	0.008	—	0.013	—	0.014	—	0.005	—
γ_{20}	0.010	—	0.008	—	0.012	—	0.012	—	0.008	—	0.008	—	0.013	—	0.014	—	0.005	—
γ_{21}	0.010	—	0.008	0.001	0.012	—	0.012	—	0.008	—	0.008	—	0.013	0.008	0.014	—	0.005	—
γ_{22}	0.010	—	0.008	—	0.012	—	0.012	—	0.008	—	0.008	—	0.013	—	0.014	—	0.005	—

This table reports the HAR coefficients (as implied by (1.3)) and the lasso coefficients. Coefficients set to 0 by the lasso procedure are indicated by dash. We only report the coefficients up to lag x_{t-22}. Lasso coefficient estimates of lags higher than x_{t-22} are reported graphically in Fig. 1.2

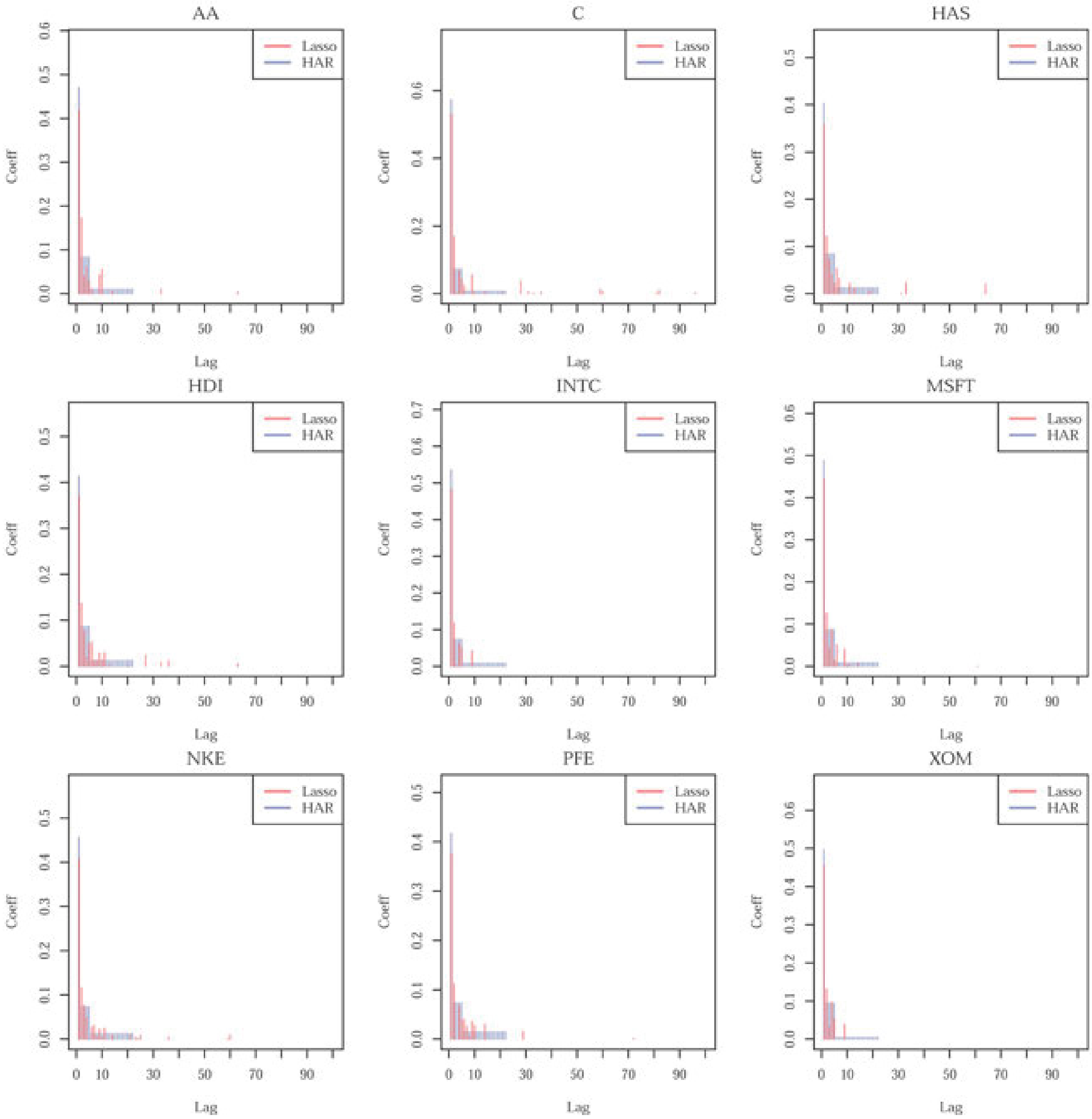
Monte Carlo Study

To assess the model selection consistency of the lasso in the case of the HAR model in finite sample we include a Monte Carlo simulation in this section. Since the lasso's model selection results depend on the signal-to-noise ratio (Bühlmann & Van De Geer 2011), it is important to have a comparable setting to assess the finite sample performance of the lasso as a model selection device. We conducted the Monte Carlo study under the assumption that the HAR model was true, in order to answer the question how effective the lasso *would* be if the HAR model *were* true. To this end, we proceeded as follows in a parametric bootstrap manner:

1. For asset $j = 1, \ldots, 9$ estimate the HAR model on the full sample of 2483 data points, which includes
   (a) Obtain c, $\hat{\beta}^{(d)}$, $\hat{\beta}^{(m)}$, $\hat{\beta}^{(w)}$ and compute $\widehat{\text{Var}}(\epsilon_t)$ as well as the derived estimates $\hat{\phi}_1^{(HAR)}, \ldots, \hat{\phi}_{22}^{(HAR)}$ via (1.3).
   (b) Compute the unconditional mean $\hat{\mu}$ (as $\hat{\gamma}_0/(1 - \sum_{i=1}^{22} \hat{\phi}_i)$) and the unconditional variance $\hat{\sigma}$ (as $\widehat{\text{Var}}(\epsilon_t)/(1 - \sum_{i=1}^{22} \hat{\phi}_i \hat{\gamma}_i)$ where $\hat{\gamma}_i$ is the autocovariance at lag i, see Brockwell & Davis (1986))
2. Resample the HAR model.
   (a) Sample $x_1, \ldots, x_{22}$ from the stationary distribution $\mathcal{N}(\hat{\mu}, \hat{\sigma})$
   (b) Compute $x_{23}, \ldots, x_{2483}$ recursively based on (1.3).
   (c) Apply the lasso as specified in Section 1.3.2 and record the lasso estimates

Step 2 is repeated 1,000 times and the results are reported in Tab. 1.3.The results clearly indicate that the HAR structure is well recovered by the lasso in this synthetic HAR setting. Although small coefficients (the monthly coefficients) are selected less often, the daily and weekly coefficients are almost always estimated to be non-zero and thus considered active.

Figure 1.2: HAR versus lasso coefficients with all predictors



Note at this point that there is indeed some contradiction with what has been reported in Tab. 1.2: The lasso does not select $\gamma_1, \ldots, \gamma_5$ for all assets and selection of lags beyond 22 is rare.[7]

We thus conclude from this Monte-Carlo application that indeed the lasso does recover the HAR model reasonably well *if* it is the true model, i.e., if we simulate from this DGP.

Rolling Window

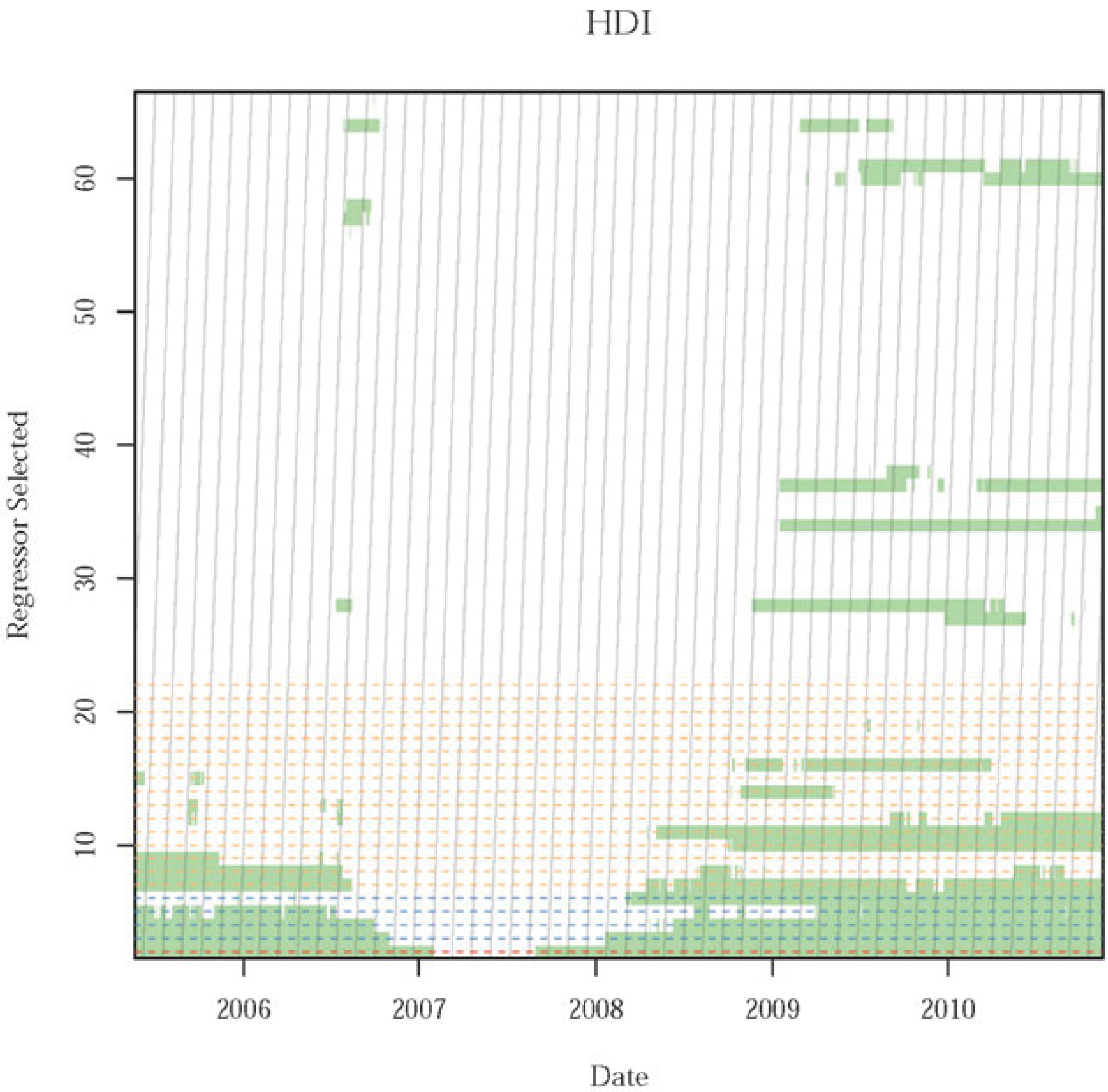
To address the question whether all of the observed in-sample selected regressors are constant over time we apply the lasso procedure in a rolling window manner. We stack our data for each asset as follows

$$X = \begin{bmatrix} x_{101} & x_{100} & \ldots & x_1 \\ x_{102} & x_{101} & \ldots & x_2 \\ \vdots & \vdots & & \vdots \\ x_n & x_{n-1} & \ldots & x_{n-100} \end{bmatrix}$$

We then estimate the lasso on the first 1,000 rows of X and roll this window of length 1,000 down to the last row of X. Pursuing this procedure we obtain 1,384 lasso estimates and record them. Fig. 1.3 contains this analysis for Citigroup, Inc. The abscissa reports the last date of the current window (the first window thus corresponds to the date of x_{1000} which in this case is May 19, 2005 and continues through Nov 15, 2010), the ordinate indicates whether or not a regressor was selected (estimated to be different from zero).

[7] Based on the percentage of times recovered we may conclude that for instance lag x_{t-15} is non-active across all nine assets (as found in Tab. 1.2) has a chance of occurring of 6.7% based on the occurrences in Tab. 1.3.

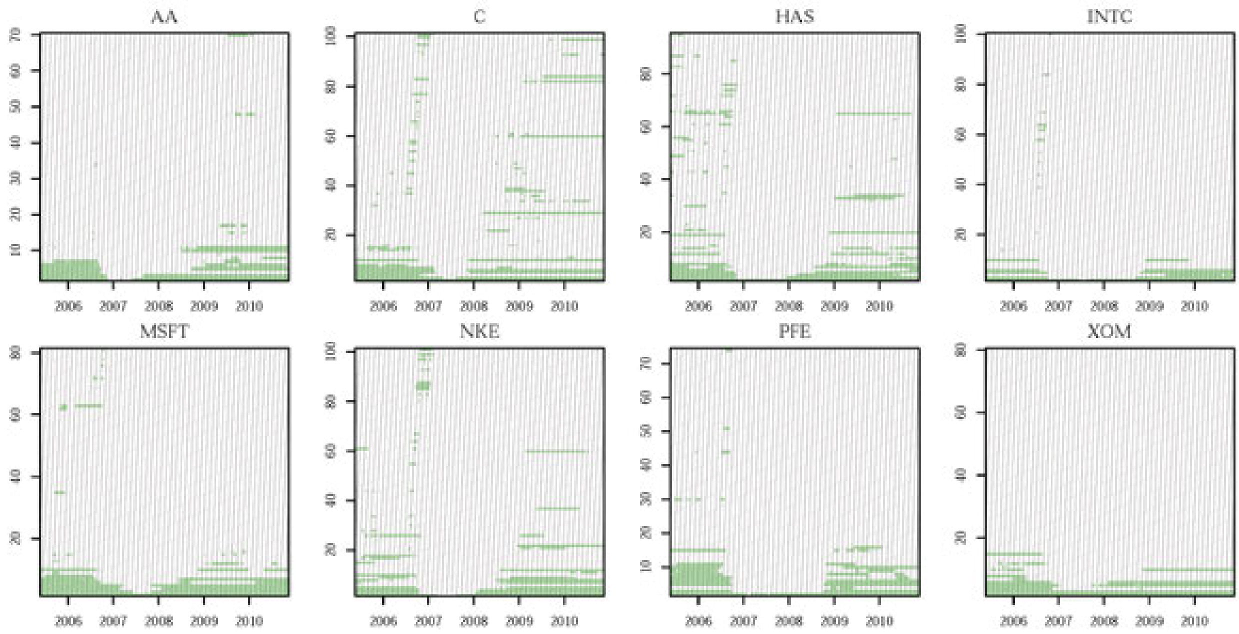
Figure 1.3: Stability of selected regressors for HDI



Stability of lasso selected regressors for Home Depot, Inc. Diagonal gray lines have slope 1, i.e., if a regressor moves along these lines then its effect is lagged by one day as the rolling window proceeds by one row (1 day)

Groups of regressors moving along the diagonal lines are likely to be noise (they are one-off events that move through the sampling window). It is also apparent from Fig. 1.3 that there is a clear break in structure during the financial crisis. The only lag which is selected

Figure 1.4: Stability of Lasso selected regressors for all assets



during the crisis is the x_{t-1} indicating that the variance process prevailing in the data is actually an AR(1)-process.

Fig. 1.4 draws the same picture for the remaining eight assets. Although there are minor differences among assets we observe a clear pattern of a "dependence breakdown" during the financial crisis. Most assets indeed also have components that can be explained by one-off events, however, we also find for HAS, HDI, and C lags that constantly get selected and remain (beyond the training window length of 1,000 observations). This may be an indication of longer-range dependence that warrants further research. As can clearly be inferred from Fig. 1.4 the dependence breakdown during the financial crisis is for some assets even more pronounced than it is for Citigroup, Inc. For these, the optimal lag structure as chosen by the lasso, sometimes reduces to a constant (e.g., HDF in Fig. 1.4). Also, there are assets that exhibit a dependence structure (i.e., by lags beyond x_{t-22}) which is not accounted for by the HAR model.

1.3.3 Out-of-sample prediction

So far we have only considered the lasso results in-sample. But the HAR has also garnered praise for its out-of-sample prediction. In a next step we thus compare the HAR's and the lasso's out-of-sample performance. We estimate the HAR model with data up to time t and compute an estimate for $t+1$ which is labeled $\widehat{\log \mathrm{RV}}_{t+1|t}^{(HAR)}$. We do the same for the lasso to obtain $\widehat{\log \mathrm{RV}}_{t+1|t}^{(lasso)}$. We proceed again in a rolling window manner but also vary the training window length (the length on which we estimate the lasso and the HAR model). To render the results comparable we report the out-of-sample prediction for different training window length but the same evaluation window (from May 12, 2009 to Nov 15, 2010 as implied by the longest training window length and resulting in 383 observations) in Tab. 1.4. To have an objective comparison we also include the random walk in our analysis. Although there is theoretical guidance for choosing λ in (1.6) we pursue a different approach. The theoretical guidance is optimal in the sense of asymptotic model selection consistency; however, this is not necessarily the best penalty for prediction. Thus, we employ the common approach of estimating the expected prediction error using cross validation.

Cross-validation in the cross sectional case is a statistically sound way of estimating the expected out-of-sample prediction error and thus determining the optimal penalty parameter (Arlot & Celisse 2010, Hastie et al. 2009). Although cross-validation (typically K-fold) is often used in practice to determine the optimal penalty parameter in a penalized regression setting (for instance in Nardi & Rinaldo (2011) and Park & Sakaori (2013)) we adopt the view of Bergmeir & Benítez (2012) and use blocked cross validation[8] to account for the time series nature of the data. When comparing the estimates of $\hat{\lambda}_{\text{opt}}$ obtained by using the regular K-fold cross validation ($\hat{\lambda}_{\text{opt}}^{(R)}$) to the estimates obtained used a K-fold blocked cross-validation ($\hat{\lambda}_{\text{opt}}^{(B)}$), we observed that $\hat{\lambda}_{\text{opt}}^{(R)} < \hat{\lambda}_{\text{opt}}^{(B)}$. From a conceptual point of view, this observation is in accordance with the result that for kernel regression the bandwidth is smaller for positively correlated errors when compared to uncorrelated errors (Hart & Wehrly 1986). Even if kernel regression and the lasso may at first appear

[8] Instead of building K blocks by randomly assigning any number in $\{1, \ldots, K\}$ to each observation and collecting the observations having the same number we use blocks with contiguous observations, such that the blocks are $\{1, \ldots, K\}, \{K+1, \ldots, 2K\}, \ldots, \{(\lfloor n/K \rfloor - 1)K + 1, \ldots, n\}$

as different approaches they can be related, exploiting the linearity of both approaches, by looking at the trace of their smoother matrix (the generalized cross-validation, GCV) which again is an estimate of the prediction error (Hastie et al. 2009).

Summarizing, we use blocked cross validation for both, empirically and theoretically founded reasons, to obtain an optimal λ in our out-of-sample procedure. We use 10 blocks to find an estimate of the optimal λ.

Table 1.4: Out-of-sample comparison

	200			400			1,000			2,000		
Asset	RW	HAR	lasso	RW	HAR	lasso	RW	HAR	lasso	RW	HAR	lasso
AA	0.160	0.129	0.142	0.160	0.126	0.126	0.160	0.125	0.123	0.160	0.124	0.121
C	0.132	0.115	0.127	0.132	0.115	0.120	0.132	0.115	0.119	0.132	0.116	0.116
HAS	0.240	0.201	0.219	0.240	0.197	0.204	0.240	0.193	0.197	0.240	0.197	0.200
HDI	0.231	0.184	0.207	0.231	0.181	0.186	0.231	0.179	0.181	0.231	0.179	0.178
INTC	0.113	0.094	0.100	0.113	0.091	0.091	0.113	0.089	0.088	0.113	0.089	0.087
MSFT	0.153	0.128	0.137	0.153	0.125	0.127	0.153	0.123	0.124	0.153	0.123	0.121
NKE	0.176	0.144	0.158	0.176	0.142	0.146	0.176	0.139	0.140	0.176	0.138	0.140
PFE	0.130	0.107	0.112	0.130	0.104	0.105	0.130	0.102	0.101	0.130	0.102	0.099
XOM	0.221	0.182	0.192	0.221	0.179	0.179	0.221	0.178	0.175	0.221	0.176	0.174

MSPE for all nine assets across training window length of 200, 400, 1,000, and 2,000 observations (rolling window). In addition to the lasso and the HAR the random walk (RW) is included.

We measure the out-of-sample performance using the mean squared prediction error (MSPE) which is computed as $\text{MSPE} = \frac{1}{n}\sum_{t=1}^{n}(\widehat{\log \text{RV}}_{t+1|t} - \log \text{RV}_{t+1})^2$ where $\widehat{\log \text{RV}}_{t+1|t}$ is the prediction obtained by either the HAR model or the lasso and n is the total number of out-of-sample predictions. Tab. 1.4 shows two points: First, both the lasso and the HAR need a certain window length to attain reasonably low mean squared prediction errors (MSPEs), although the HAR model is markedly better for small training window sizes. Second, for longer training windows, the lasso and the HAR are almost equal in terms of MSPE.

To better understand these results we further report the evaluation over different out-of-sample periods: Pre-crisis, post-crisis, and full sample. The date for the beginning of the financial crisis was set to Sep 1, 2007. For the relevant training window lengths (i.e., 1,000 days and 2,000 days) we kept the maximal out-of-sample period which, unlike Tab. 1.4, results in evaluation windows of different lengths. The difference in MSPE is then tested using the Diebold-Mariano test (Diebold & Mariano 1995). These results are reported in Tab. 1.5.

Table 1.5: Diebold-Mariano (Diebold & Mariano 1995) tests of equal predictive ability

			AA	C	HAS	HDI	INTC	MSFT	NKE	PFE	XOM
1,000	Total	Mean Diff.	0.002	-0.001	0.001	-0.003	-0.001	-0.002	-0.006	**-0.007**	-0.001
	(n=1'383)	*p*-value	*0.38*	*0.86*	*0.85*	*0.34*	*0.49*	*0.35*	*0.18*	*0.00*	*0.48*
	PreCrisis	Mean Diff.	0.002	0.000	0.002	-0.006	0.001	-0.001	-0.012	**-0.008**	0.000
	(n=575)	*p*-value	*0.57*	*1.00*	*0.78*	*0.22*	*0.50*	*0.55*	*0.17*	*0.01*	*1.00*
	PostCrisis	Mean Diff.	0.002	-0.001	0.000	0.000	-0.003	-0.003	-0.001	**-0.006**	-0.003
	(n=808)	*p*-value	*0.49*	*0.84*	*0.96*	*0.94*	*0.26*	*0.44*	*0.69*	*0.05*	*0.39*
2,000	Total	Mean Diff.	0.002	0.000	-0.003	0.001	**0.002**	0.001	-0.002	0.003	0.002
	(n=383)	*p*-value	*0.38*	*0.92*	*0.44*	*0.72*	*0.09*	*0.62*	*0.30*	*0.18*	*0.33*
	PreCrisis	Mean Diff.	—	—	—	—	—	—	—	—	—
		p-value	—	—	—	—	—	—	—	—	—
	PostCrisis	Mean Diff.	0.002	0.000	-0.003	0.001	**0.002**	0.001	-0.002	0.003	0.002
	(n=383)	*p*-value	*0.38*	*0.91*	*0.44*	*0.72*	*0.09*	*0.62*	*0.29*	*0.18*	*0.33*

Difference in MSPE ($\text{MSPE}_{\text{HAR}} - \text{MSPE}_{\text{lasso}}$) are reported together with p-values from the Diebold-Mariano (Newey-West (Newey & West 1987) adjusted). The differences and p-values are reported for different training windows (1,000, 2,000) and before/after the financial crisis. Differences significant at 0.1 are typeset in boldface

Although there are a small number of rejections of the null we find no consistent pattern, neither in favor of the HAR nor in favor of the lasso. Investigating the predictions $\widehat{\log \text{RV}}^{(\text{lasso})}_{t+1|t}$ and $\widehat{\log \text{RV}}^{(\text{HAR})}_{t+1|t}$ in the sense of Mincer & Zarnowitz (1969) we find no evidence of either of the models (reported in Appendix A.5) being more often unbiased.

To be retained at this stage is that there is no clear evidence that either of the two models is genuinely better suited to forecast realized variance out-of-sample.

1.3.4 Risk Management Application

To test the predictions obtained from the lasso and the HAR model from a different angle, we include a risk management application. The value-at-risk of an asset to the level α is given as

$$\text{VaR}^t_\alpha = -\inf\{x \in \mathbb{R} | \, \text{P}(X_t \leq x) \geq 1 - \alpha\} \tag{1.9}$$

where X_t is the daily log-return of an asset.[9] Under the assumption, which also underlies the computation of realized variance, that an asset's return X_t is given as[10]

$$X_t = \mu_t + \sigma_t \cdot Z_t$$

we can readily compute (assuming a scale-location family with continuous distribution function) as

$$\text{VaR}_\alpha(X) = \mu_t + \sigma_t q_{1-\alpha} \tag{1.10}$$

where $q_{1-\alpha}$ is the $1 - \alpha$ quantile of the standardized distribution Z_t, μ_t the conditional mean, and σ_t the conditional volatility of X.

As the distribution for Z_t we use the standard normal distribution as well as the empirical distribution after (quasi-)standardizing X_t with μ_t and estimates of σ_t as obtained by the RV_t estimates. Since we are aiming for a realistic benchmark we do not employ backtesting for the value-at-risk but conduct an out-of-sample analysis and predict

$$\text{VaR}_\alpha^{t+1|t} = \mu_t + \sigma_{t+1|t} q_{1-\alpha} \tag{1.11}$$

where $\sigma_{t+1|t}$ is again obtained based on $\text{RV}_{t+1|t}$ estimates by either the lasso or the HAR model.

To do so, we estimate both models on window lengths of $n = 200, 400, 1,000, 2,000$ observations to obtain a forecast $\widehat{\log \text{RV}}_{t+1|t}$. To get an optimal forecast (in the sense of Proietti & Lütkepohl (2013) and Appendix A.2) of the actual volatility we compute $\hat{\sigma}_{t+1|t}$ as

$$\hat{\sigma}_{t+1|t} = \sqrt{\exp(\widehat{\log \text{RV}}_{t+1|t} + \frac{\tilde{\sigma}^2}{2})} \tag{1.12}$$

where $\tilde{\sigma}^2$ is the variance of $\widehat{\log \text{RV}}_{t+1|t}$. The hit ratios are then defined as

$$\text{HR}_\alpha^{M\,(D)} = \frac{\#\{x_{t+1} < -\text{VaR}_\alpha^{t+1|t}\}}{n} \tag{1.13}$$

[9] We define the value-at-risk compliant to the risk management literature: Instead of working with the usual distribution, we premultipliy with −1 such that losses are positive resulting in the mnemonic that a greater VaR means greater risk.

[10] Strictly speaking the assumptions of computing realized variance also allow for jumps (depending on the estimator) to contribute to the return X_t. For reasons of simplicity, we exclude this component.

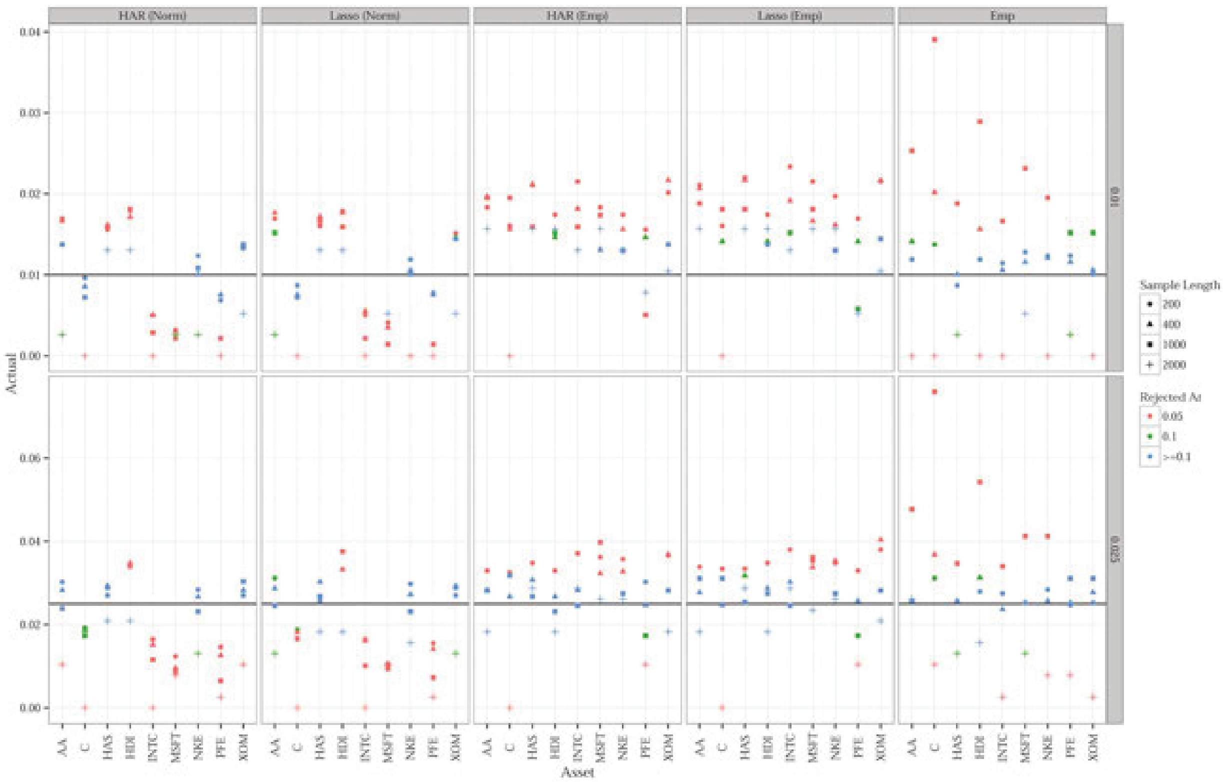
where 'M' can either be 'HAR' or 'lasso' depending on how $\sigma_{t+1|t}$ of (1.12) is computed (either by the HAR-model or our lasso approach), and 'D' is either 'Norm' or 'Emp' depending on how $q_{1-\alpha}$ in (1.11) is computed (quantiles of a $\mathcal{N}(0,1)$ distribution or quantiles of the standardized empirical distribution). In all cases we compute the conditional mean as $\mu_t = \frac{1}{n}\sum_{i=1}^{n} r_{t-n+i}$.

To contrast these estimates we also implement a naive estimator of the value-at-risk by simply taking the empirical α-quantile of the distribution of the log-returns, i.e.

$$\mathrm{HR}_{\alpha}^{Emp} = \frac{\#\{x_{t+1} < \hat{q}_{1-\alpha}\}}{n},$$

where $\hat{q}_{1-\alpha}$ is the empirical $1-\alpha$ quantile of $\{x_{t-n+1}, \ldots, x_t\}$.

Figure 1.5: Actual hit ratios.



The columns show the different estimators of HR_{α}, the rows show the levels of $\alpha = 99, 97.5\%$. The horizontal lines are the theoretical levels $(1-\alpha)$ of the VaR. The color indicates the p value of Kupiec's (1995) test against the theoretical level.

Fig. 1.5 clearly shows that there is again no systematic difference between the $\mathrm{HR}_{\alpha}^{HAR\ D}$ and $\mathrm{HR}_{\alpha}^{lasso\ D}$. Both are too aggressive (producing a VaR which is too low and thus is violated more often than theoretically specified) when the emprical distribution is used for the standardized innovations, and less so when the $\mathcal{N}(0,1)$-distribution is used for the standardized innovations. What becomes apparent from Fig. 1.5 is that the influence of the assumption on the distribution as well as the asset in question is much more crucial than the model used to forecast volatility. Compared to the simple model of estimating the VaR by simply taking the empirical quantiles the results are disappointing: There is no apparent outperformance of computing the VaR with volatility forecasts obtained by either the HAR or the lasso over the simple (but rather effective) historical quantiles over short training periods. This is all the more so, when looking at the rejections of $\mathcal{H}_0$ under Kupiec's (1995) test (assuming the correct level for the VaR). It is less often rejected for the 'Emp' than for any realized variance model at both the 5% and 10% level of Kupiec's test.

The particularly poor performance of all VaR forecasts for Citigroup, Inc. is related to the turbulent times the stock went through during the financial crisis resulting in pronounced non-normality of the $\log \mathrm{RV}_t$ as reported in Fig. 1.1 (b) as well as non-normality of the log-returns reported in Fig. A.5.

1.4 Conclusions and Further Research

We conclude that the lasso does not recover the HAR model. We consider this as evidence against the presumption that HAR model is the true DGP since, first, we have theoretically founded reason to believe that the lasso should detect the HAR model, and, second, we provided empirical evidence on synthetic data that the lasso does recover the HAR model if the data stem from this DGP.

In addition, the lasso and the HAR model appear to be indistinguishable from an out-of-sample performance point of view: Neither the HAR nor the lasso excels in an out-of-sample prediction exercise. When we look at a more economically meaningful comparison using value-at-risk prediction, both models fare equally poorly with no noticeable differences in favor of either of the two.

The argument above and the selection of only near-lags (in the whole sample, and even more pronouncedly during the crisis) leads us to the hypothesis that in fact the realized variance dynamics are much better explained by shorter horizon models. Our results are in line with empirical evidence shown in Chen, Härdle & Pigorsch (2010), eventually hinting at the possibility that the seemingly long-memory dynamics of the realized variance time series are in fact spurious. Arguments against this view are the lags which are selected and persist: This actually indicates that there might be some long range dependence which warrants further research.

We thus conclude that the HAR model may not be the true model. However, it captures – as does the lasso – a linear footprint of the possibly non-linear volatility dynamics that can be used for volatility forecasting. Given the equal out-of-sample performance of the two approaches we see potential for further research in this domain: Although adding additional predictors other than the lagged values of the realized volatilities themselves expels us from the thorough theoretical model selection framework established in this paper, we anticipate further insights with regard to e.g., volatility spillovers (including other assets, markets, etc. as predictors) or calendar effects (adding day-of-week dummies to the lasso regression).

Table 1.3: Percentage of HAR coefficients recovered

Lag	AA	C	HAS	HDI	INTC	MSFT	NKE	PFE	XOM
x_{t-1}	100	100	100	100	100	100	100	100	100
x_{t-2}	100	100	100	100	100	100	100	97	100
x_{t-3}	100	100	100	100	100	100	100	98	100
x_{t-4}	100	100	100	100	100	100	100	97	100
x_{t-5}	100	100	100	100	99	99	100	96	100
x_{t-6}	42	43	61	54	18	21	52	23	12
x_{t-7}	37	39	59	53	17	18	50	21	9
x_{t-8}	36	37	61	54	15	16	50	20	6
x_{t-9}	32	32	54	49	9	10	44	15	2
x_{t-10}	34	34	56	50	9	10	45	16	1
x_{t-11}	31	31	54	45	7	9	42	14	1
x_{t-12}	28	30	56	48	7	9	44	15	1
x_{t-13}	27	28	55	47	5	7	43	14	1
x_{t-14}	28	29	58	51	6	7	46	13	0
x_{t-15}	25	24	55	46	4	4	42	10	0
x_{t-16}	25	25	53	44	4	5	40	12	0
x_{t-17}	19	20	51	42	2	2	36	8	0
x_{t-18}	20	20	51	43	2	2	38	8	0
x_{t-19}	16	17	46	36	1	2	31	6	0
x_{t-20}	12	12	42	31	1	1	27	3	0
x_{t-21}	10	12	41	29	0	0	25	3	0
x_{t-22}	8	9	34	25	0	0	20	2	0
x_{t-23}	3	6	13	8	0	0	6	1	0
x_{t-24}	2	4	8	6	0	0	4	0	0
x_{t-25}	1	4	8	5	0	0	3	0	0
x_{t-26}	1	3	5	4	0	0	3	0	0
x_{t-27}	0	2	4	2	0	0	2	0	0
x_{t-28}	0	2	4	3	0	0	1	0	0
x_{t-29}	1	2	4	2	0	0	2	0	0
x_{t-30}	0	1	3	2	0	0	1	0	0
x_{t-31}	0	2	3	2	0	0	1	0	0
x_{t-32}	0	1	3	1	0	0	1	0	0
x_{t-33}	0	1	2	1	0	0	1	0	0
x_{t-34}	0	1	2	1	0	0	0	0	0
x_{t-35}	0	1	1	0	0	0	0	0	0
x_{t-36}	0	1	2	1	0	0	1	0	0
x_{t-37}	0	1	2	1	0	0	1	0	0
x_{t-38}	0	1	2	1	0	0	0	0	0
x_{t-39}	0	1	1	0	0	0	0	0	0
x_{t-40}	0	0	1	0	0	0	0	0	0
x_{t-41}	0	0	1	0	0	0	0	0	0
x_{t-42}	0	0	1	0	0	0	0	0	0
x_{t-43}	0	0	1	0	0	0	0	0	0
x_{t-44}	0	1	1	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_{t-100}	0	0	0	0	0	0	0	0	0

Number of times (out 1,000 replications) a lag has been selected (estimated as non-zero) by the lasso in percent. Omitted rows contain zero only.

Chapter 2

Markets from East to West, News and Volatility Spillovers: Comparing Forecast Accuracy

Simon D. Knaus

Abstract

This paper investigates the role of volatility spillovers, macroeconomic news, intra-week seasonality and the leverage effect for realized volatility models. To understand the role of news existing models, augmented by this new combined information set, are revisited. The least absolute shrinkage and selection operator is employed in a vector autoregressive setting to assess the relevance of volatility spillovers to modeling the S&P 500's realized volatility. A combined model is then proposed which features a data-driven selection of regressors to include in a model of realized volatility. These models are compared in a strict out-of-sample prediction comparison together with a value-at-risk application. A superior performance of models augmented with this new information set is witnessed in the prediction comparison. Compared to existing models, a considerably shorter lag structure delivers already good forecasting performance.

JEL: C58, C63, C49

Keywords: Realized Volatility, Heterogeneous Autoregressive Model, Lasso, Spillover, VAR

2.1 Introduction

Contrary to asset returns, it is widely accepted that the volatility of asset returns is indeed predictable to a certain extent. Apart from scientific interest per se in this topic, accurate forecasts of volatility are also paramount in many applications in finance ranging from asset allocation to risk management.

The introduction of the autoregressive conditional heteroscedastic (ARCH) model by Engle (1982) and the extension to the generalized ARCH (GARCH) model by Bollerslev (1986) constitute seminal contributions to the realm of volatility modeling. The GARCH model provides a framework in which to model conditional daily volatility and is able to replicate empirically observed facts such as volatility clustering. Countless extensions of the GARCH model have found their way to the literature. These extensions feature asymmetries, spillovers, and similar phenomena. While all these models help the understanding of volatility, volatility as such is still unobserved and model-dependent.

A considerable advancement in this field resulted from the advent of high-frequency data with frequencies as high as tick-by-tick in recent years. This abundance of data led to a new strand of research addressing the actual estimation of daily volatility using *intra-day* observations of asset returns. This line of research renders thus far unobservable volatility observable by asymptotic arguments. Much of it is indebted to seminal contributions of Bollerslev & Wright (2001), Barndorff-Nielsen & Shephard (2002*a*), and Andersen, Bollerslev, Diebold & Labys (2003), among others. Hence, one can obtain a value of the daily volatility which – until then – could only be estimated from daily data with a specific model.

This estimate of daily volatility starts from the assumption that the log price process p_t is a (jump-) diffusion, i.e., $p_t = \mu_t \mathrm{d}t + \sigma_t \mathrm{d}W$ where μ_t and σ_t are the instantaneous mean and volatility, and W_t a Brownian motion. The daily log return (with a day normalized to 1) is given as $r_t = p_t - p_{t-1}$ which can be shown to be $\mathcal{N}(\int_{t-1}^{t} \mu_s \mathrm{d}s, \int_{t-1}^{t} \sigma_s^2 \mathrm{d}s)$ distributed (McAleer & Medeiros 2008). The term $\int_{t-1}^{t} \sigma_s^2 \mathrm{d}s$ is referred to as *integrated variance* (IV_t) and can be estimated by summing the squared intra-day returns. Realized volatility,

$\mathrm{RV}_t = \sum_{j=1}^{N}(p_{t_j} - p_{t_{j-1}})^2$, converges uniformly in probability to IV_t as the grid becomes finer (a more rigorous discussion is provided in the appendix). This measure RV_t of return variability is commonly called, by slight abuse of language, *realized volatility*. Its square root is thus an estimate of daily volatility.

With a series of daily realized volatilities at hand, the question of modeling can be approached. The series of realized volatilities inherits many stylized facts also found in earlier works on conditional volatility time series. A well-documented empirical fact is the pronounced long-range dependence and near log-normality of the unconditional distribution. These observations have led to two fundamental classes of models: first, a fractionally integrated autoregressive moving average model (ARFIMA) pioneered by Andersen et al. (2003), and, second, a heterogeneous autoregressive model (HAR) advocated by Corsi (2009). While the latter class of models is not able to reproduce long-memory time series formally, it still enjoys great popularity, mostly due to its good empirical performance, ease of implementation, and economic interpretability. More recent research is embarking on new ideas to address the specificities of the realized volatility time series which tries to explain the long memory behavior in terms of multiple regimes that may only prevail over a certain time (Scharth & Medeiros 2009, Chen et al. 2010). Most of these models still have an autoregressive (AR) component and thus to a certain extent favor the (H)AR view.

A natural next step is to investigate these models and their ability to incorporate well-documented facts already known from the earlier GARCH period: among these are, the leverage effect which postulates that negative returns entail a greater volatility, the effect of news, intra-week seasonality, and volatility spillovers. The benefit of including macroeconomic news has for instance been addressed by Martens, van Dijk & de Pooter (2009) and is also employed by Scharth & Medeiros (2009). A recent study addressing the question of volatility spillovers for realized volatilities is found in Dimpfl & Jung's (2012) work.

This paper contributes to these lines of research. Combining both, volatility spillovers and news (in the wide sense, including intra-week seasonality, macroeconomic news, and leverage effect), the role of each of these components is investigated. Based on these results, the out-of-sample performance, the overarching goal, is analyzed to understand the benefits of this augmented information set. To achieve this, a novel way to assess the importance of volatility spillovers, using the least absolute selection and shrinkage operator (lasso) in a vector autoregressive (VAR) context, is proposed. To address the relevance of the information set augmented by news, existing models are revisited. The lasso is also employed in univariate volatility modeling to corroborate the findings of earlier models. Hence, the present work also touches on the question of whether volatility can be modeled in a self-sufficient way or whether the inclusion of exogenous factors is required. To address the question of the results' robustness with regard to the measurement of realized volatility, all the analyses are conducted with different estimators for realized volatility. The entire paper primarily uses the bipower variation (Barndorff-Nielsen & Shephard 2004), however, the same analyses are also carried out using the naive estimator for the realized volatility as introduced earlier. These results are deferred to the appendix together with a complete discussion of the estimators and data used. To assess the final models' performance realistically, I adhere to a strict separation of the in-sample period (training period) and the out-of-sample period (evaluation period). Hence, explorative investigations are only carried out using the training period while the final models are tested on the evaluation period.

The rest of the paper is structured as follows: The first section revisits the idea of Dimpfl & Jung (2012) and casts it in a lasso framework. The second section reviews the role of news, including the day-of-the-week and the leverage effect, on volatility modeling and adds a model selection perspective extending Audrino & Knaus (2012). The third section, collecting the findings from the previous two sections, proposes several models which are then assessed in an out-of-sample forecasting comparison featuring pure volatility forecasts as well as more practically oriented value-at-risk forecasts. The last section discusses the results and concludes. A detailed description of the data can be found in the appendix.

2.2 Volatility Spillovers

2.2.1 Overview

The effect of one asset on another asset's return and volatility is a widely studied subject. This research dates back at least to Hamao, Masulis & Ng (1990), and earlier to Hilliard (1979) who focuses on asset prices. Although the subject has been studied for quite some time there is no – to my knowledge – commonly accepted taxonomy: "spillover", "contagion", "comovement", or simply dependence are often used interchangeably to capture the fact that there may be an effect of one asset's volatility on another asset's volatility.

I adopt the view that the ultimate goal of volatility models is prediction. Spillover can then be seen in the sense of Granger's idea of causality: "We say that Y_t is causing X_t if we are better able to predict X_t using all available information than if the information apart from Y_t had been used" Granger (1969, Definition 1). In the present paper I use the term spillover in the above sense, i.e., there is spillover, if one is better able to predict one variable by including another variable.

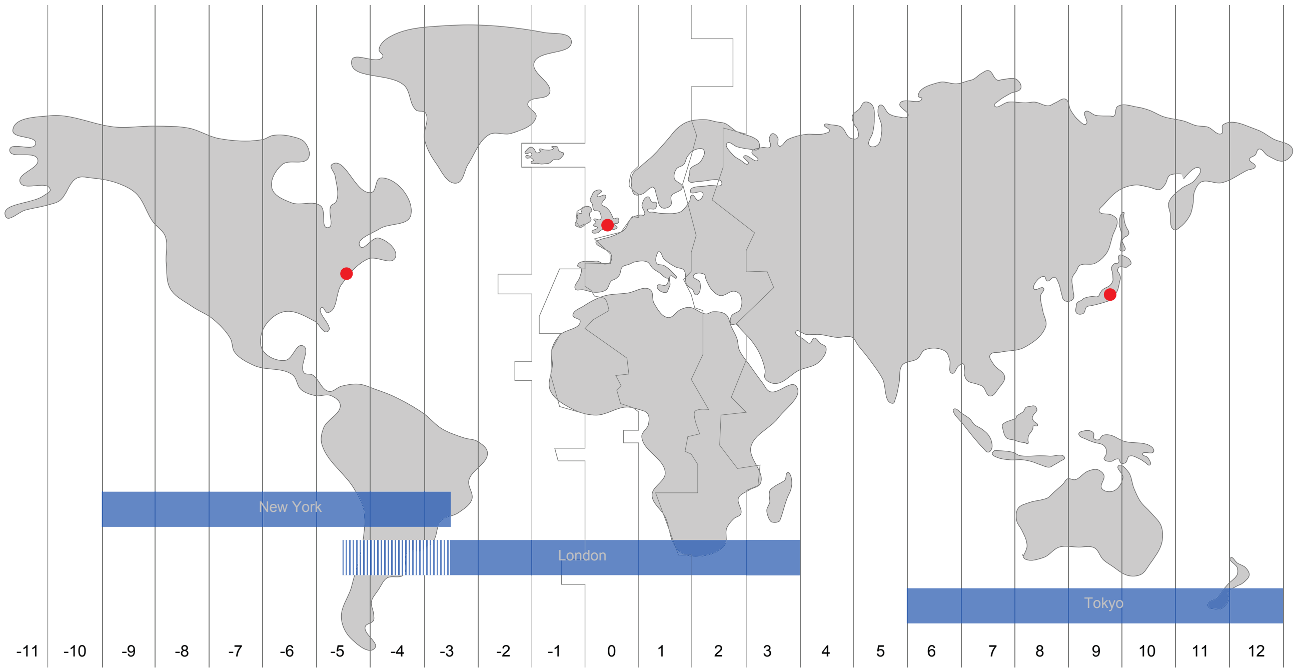
The research on volatility spillover is ample. A very recent overview of spillover analyses in Asian markets is provided in Engle, Gallo & Velucchi (2012). An approach measuring spillovers is developed in Diebold & Yilmaz (2009) and extended in Diebold & Yilmaz (2012). Other recent studies that investigate volatility spillovers include: Savva, Osborn & Gill (2009), who consider the integration of the German and the French market before and after the introduction of the Euro; Gębka & Serwa (2007), who look at intra- and inter-regional spillovers in emerging markets; Golosnoy, Gribisch & Liesenfeld (2012), who propose a three-phase intraday model to analyze spillover between the US and the German market; Jiang, Konstantinidi & Skiadopoulos (2012), who consider the effect of news announcements and the subsequent volatility spillovers by investigating option volatility.

Clearly, volatility spillover constitute a very active research field. However, most work focuses on daily data. The most influential paper on this essays' view is Dimpfl & Jung (2012), which uses intraday data to compute realized volatility. Dimpfl & Jung study the effects of volatility transmission with a multivariate model measuring realized volatility itself univariately. Their view on volatility spillover modeling is captured by a VAR process that attempts to explain the volatility series using a daily, a weekly and a semi-annual component. This choice of components is inspired by Corsi's (2009) heterogeneous autoregressive model (HAR) which originally suggested the use of daily, weekly, and monthly linear components to model realized volatility univariately.

Extending the work of Dimpfl & Jung (2012), I relax the assumptions on the component structure. To this end, realized volatility is computed univariately and the specific temporal structure is exploited in a structural VAR setting.

With the goal of modeling the S&P 500's volatility, the current work limits the attention to three major stock markets represented by the most important cash indices: the S&P 500 in New York (SP), the FTSE 100 in London (FT), and the Nikkei 225 in Japan (NE). A complete description of the data is available in the appendix. The trading hours and geographical situation of these indices are presented in Fig. 2.1.

Figure 2.1: Index trading hours



Schematic illustration of time zones and market opening. The three markets have the following opening hours: Nikkei from 9.00 a.m. to 3.00 p.m (0h00 to 06h00 UTC), FTSE from 8.00 a.m. to 4.30 p.m (08h00 to 16h30 UTC), and S&P 500 from 9.30 a.m. to 4.00 p.m (14h30 to 21h00 UTC). Note that London and New York have daylight saving time such that the opening hours are 07h00 to 15h30 UTC+1 and 13h30 to 20h00 UTC+1 respectively during the summer time. The situation illustrated in Fig. 2.1 shows the situation during daylight saving time.

As can be observed, there is no temporal overlap between Tokyo and London, whereas London and New York have concurrent trading from 14h30 to 16h30 UTC. Since the ultimate goal is to devise a model for the SP's realized volatility[1], I follow the common approach of truncating the observations of the FT to the opening of the SP. This is also referred to as *pseudo closing time* (Savva et al. 2009, Dimpfl & Jung 2012). Following this argument, one can model the realized volatility analogously to Dimpfl & Jung (2012) as described in more detail in the next paragraph.

[1]The current approach is by no means limited to modeling the SP's realized volatility. Instead of following a calendar date definition of a day, one might as well work with fictitious days such that the FT and SP precede the NE and the NE's realized volatility is the ultimate goal of prediction.

2.2.2 Theoretical Set-Up

Let $y_{i,t}$ be the log realized volatility, observed for market $i = 1$ (NE), $i = 2$ (FT), and $i = 3$ (SP) where the realized volatility is computed as outlined in the appendix, i.e., $y_{i,t} = \log \mathrm{RV}_{i,t}$. The realized volatility of the FT is only computed up to the opening time of the SP.[2]

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ b_{21,0} & 0 & 0 \\ b_{31,0} & b_{32,0} & 0 \end{pmatrix} \begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix}$$
$$+ \sum_{l=1}^{p} \begin{pmatrix} b_{11,l} & b_{12,l} & b_{13,l} \\ b_{21,l} & b_{22,l} & b_{23,l} \\ b_{31,l} & b_{32,l} & b_{33,l} \end{pmatrix} \begin{pmatrix} y_{1,t-l} \\ y_{2,t-l} \\ y_{3,t-l} \end{pmatrix} + \begin{pmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \end{pmatrix}$$

or, more compactly,

$$y_t = a + B_0 y_t + \sum_{l=1}^{p} B_l y_{t-l} + u_t. \tag{2.1}$$

The specific form of (2.1) follows from the temporal structure shown in Fig. 2.1 and is identical to that of Dimpfl & Jung (2012). Dimpfl & Jung use restrictions of the form

$$B_l = \begin{cases} C_l^{(d)} + \frac{1}{5} C_l^{(w)} + \frac{1}{120} C_l^{(sa)} & \text{for } l = 1 \\ \frac{1}{5} C_l^{(w)} + \frac{1}{120} C_l^{(sa)} & \text{for } l = 2, \dots, 5 \\ \frac{1}{120} C_l^{(sa)} & \text{for } l = 6, \dots, 120 \end{cases}$$

such that the model is an autoregression on lagged daily, average weekly, and average semi-annual volatility. Unlike Dimpfl & Jung (2012), I do not restrict the autoregressive parameters to be specific aggregation frequencies. I will determine these from the data as will be discussed in the following paragraph.

[2] For an illustration showing the relation between the pseudo-closing realized volatility versus the all day realized volatility refer to Fig. B.3

Note at this stage that if y_t is a demeaned series ($a_i = 0 \; \forall i$) and $C = (I - B_0)^{-1}$ one can write (2.1) as

$$y_t = \sum_{l=1}^{p} CB_l y_{t-l} + Cu_t = \sum_{l=1}^{p} \tilde{B}_l y_{t-l} + \tilde{u}_t. \tag{2.2}$$

Given the specific form of B_0, the structural VAR is just identified (Hamilton 1994), hence one can estimate (2.2) by means of OLS. In the subsequent paragraph I follow Kock & Callot (2012*a*) in the exposition of the adaptive lasso and its applications to vector autoregressive processes as given in (2.2).

2.2.3 Lasso in the VAR Structure

The least absolute shrinkage and selection operator (lasso) was originally introduced by Tibshirani (1996) to address regression-linked problems in the cross-sectional case. The lasso is not only a robust regression device (by using shrinkage, collinear or near-collinear predictors can still be used in a linear framework) but also a model selection operator. The original lasso can be used to tackle cross-sectional models of the form $y_i = \sum_{j=1}^{p} \beta_i x_{j,i} + \epsilon_i$, or more compactly, $y = X\beta + \epsilon$ where X is the matrix of predictors, β the (true) $p \times 1$ parameter vector (with potentially some $\beta_i = 0$) and ϵ the error term. Typically it is assumed that the regressors as well as the response variable are standardized; otherwise prior normalization with respect to the Mahalanobis distance is required.

The lasso estimator of β, $\hat{\beta}_{\text{lasso},\lambda}$ then minimizes

$$L(\beta) = ||y - X\beta||_2^2 + \lambda \sum_{i=1}^{p} |\beta_i|.$$

for each λ such that a $\hat{\beta}_{\text{lasso},\lambda}$ sequence is produced. The combination of the L^2 (the objective) and the L^1 (the penalty) norm makes the lasso a computationally efficient model selection device (Hastie et al. 2009). It can be shown that the lasso recovers the true active parameters $\mathcal{A} = \{i|\beta_i \neq 0\}$ with probability 1 asymptotically for an appropriate choice of λ (Bühlmann & Van De Geer 2011).

An important extension of the lasso was contributed by Zou (2006) who proposes the *adaptive* lasso estimator which minimizes

$$L(\beta) = \|y - X\beta\|_2^2 + \lambda \sum_{i=1}^{p} \xi_i |\beta_i|$$

where ξ_i are specific weights penalizing the regressors more or less severely. A large ξ_i weighs a specific coefficient β_i down whereas a small ξ_i puts a smaller penalty on β_i. The standard lasso is thus a special case of the adaptive lasso with $\xi_i = 1$ for all $i = 1, \ldots, p$. The adaptive variant of the lasso has the advantage that it relaxes the conditions under which the correct sparsity pattern $\mathcal{A}$ is identified (Bühlmann & Van De Geer 2011) and thus enjoys great popularity since it is easy to implement, remains computationally efficient, and renders model selection (within the class of linear models) feasible on a theoretically grounded footing.

Building on this strand of literature, the lasso has recently also been extended to time series regression, namely, to autoregressive processes. An important contribution in this field has been made by Nardi & Rinaldo (2011), who establish that the lasso is also able to recover the true lag structure of a univariate AR process. An even more important contribution is the extension of the adaptive lasso to vector autoregressions (Kock & Callot 2012*a*, Kock & Callot 2012*b*).

It is the results of Kock & Callot (2012*a*) that will be of integral importance to the current application. Hence, in the following I adopt the notation and terminology of Kock & Callot in the exposition of the problem.

Observe that (2.2) can be written as

$$y = X\beta + \tilde{u} \tag{2.3}$$

where $Z_t = (y'_{t-1}, \ldots, y'_{t-p})$, $Z = (Z_T, \ldots, Z_1)'$, $X = I_N \otimes Z$, and $y = (y'_1, \ldots, y_N)'$ such that β contains the N^2p parameters of $\tilde{B}_l$, $l = 1, \ldots, p$ and $\tilde{u} = (\tilde{u}'_1, \ldots, \tilde{u}'_N)'$. The parameters $\tilde{B}_l$

can be recovered by the identity $\beta = \text{vec}((\tilde{B}_1, \ldots, \tilde{B}_p)')$ where vec is the usual columnwise vectorization operator.

The adaptive lasso estimator for (2.3) minimizes

$$L_T(\beta) = ||y - X\beta||_2^2 + \lambda_T \sum_{i=1}^{N^2 p} \xi_i |\beta_i| \tag{2.4}$$

where ξ_i are weights that are commonly chosen as $\xi_i = |\hat{\beta}_{OLS,i}|^{-\gamma}$ where $\hat{\beta}_{OLS}$ is an OLS estimate for β in (2.3) and γ is commonly assumed to be 1.

Kock & Callot (2012*a*) then proceed and show that if (2.2) is a stationary VAR with

(i) $\tilde{u}_{i,t}$ have a finite fourth moments for all $i = 1, \ldots, N$ and that $\tilde{u}_t$ is zero mean i.i.d. distributed random vectors with covariance matrix Σ.

(ii) $\mathbb{E}(\frac{1}{T} Z'Z)$ is positive definite

then the adaptive lasso is *oracle efficient* if $\lambda_T / \sqrt{T} \to 0$ and $\lambda_T T^{-1/2+\gamma/2} \to \infty$ additionally holds. The use of assumption (i) is standard and ensures convergence to the limiting distribution and assumption (ii) rules out collinearity of the lagged variables. For more details see Kock & Callot (2012*a*).

Oracle efficiency is captured by the following three points (Kock & Callot 2012*a*, Theorem 1):

(i) $|| \sqrt{T}(\hat{\beta} - \beta)||_2^2 \in O_{\text{P}}(1)$,

(ii) $\text{P}(\hat{\beta}_{\mathcal{A}^c} = 0) \to 1$,

(iii) $\sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A}})$ is multivariately normally distributed with mean zero.

$\mathcal{A}$ is again the set of active coefficients, and $\mathcal{A}^c$ is the set of inactive coefficients ($\mathcal{A}^c = \{1, \ldots, N^2 p\} \setminus \mathcal{A}$).

The result of Kock & Callot (2012*a*) thus enables one to use the lasso in the VAR setting. Asymptotically the true sparsity pattern will be detected by the lasso and consequently, some entries in $\tilde{B}_i$ for $i = 1, \ldots, p$ will be set to zero. In other words, the true non-active predictors in (2.3) will be set to zero as the number of observations increases. Although this only holds asymptotically, there is evidence that the adaptive lasso applied to the VAR model also delivers the expected results in finite sample. Kock & Callot (2012*b*) provide Monte Carlo evidence for the detection of the correct sparsity pattern as well as a real-world application in Kock & Callot (2012*a*) whereby the adaptive VAR lasso is successfully employed to forecast macroeconomic time series.

Once (2.2) is estimated by the lasso in the spirit of (2.3) one can proceed to recover the structural parameters of (2.1).[3] Hence, the theoretical framework required to determine the predictors and their relevant lags of (2.1) in a data-driven manner is established. Adhering to the convention of standardized variables, the magnitude of the coefficients can then be seen as importance to reduce the L^2 loss in (2.4). By the inversely monotonic relationship of the coefficient of determination to the L^2 loss, this can thus also be seen as the importance to increase R^2.

[3] If the usual assumption

$$\mathbb{E}(u_t u'_{t+j}) = \begin{cases} D & \text{if } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

holds for (2.1) with a diagonal matrix D we have for $\tilde{u}_t$ of (2.2)

$$\mathbb{E}(\tilde{u}_t \tilde{u}'_{t+j}) = \begin{cases} CDC' & \text{if } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Since the system is just identified, the matrix C (Hamilton 1994, 11.6) can be recovered using the the unique triangular factorization of any positive definite matrix Ω into $L\tilde{\Delta}L'$ where L is a lower triangular matrix with ones on the diagonal and $\tilde{\Delta}$ a diagonal matrix. This unique $L\tilde{\Delta}L'$ decomposition follows from the Cholesky decomposition of Ω into UU' with a unique upper triangular matrix U. Let $U = [u_{ij}]_{\{1 \le i,j \le p\}}$ and $\Delta = [\delta_{ij}]_{\{1 \le i,j \le p\}}$ with $\delta_{jj} = u_{jj}$ and $\delta_{ij} = 0$ for $i \neq j$. Then $L = U'\Delta^{-1}$ and $\Omega = L\Delta^2 L' = L\tilde{\Delta}L'$.

Since C is precisely a lower triangular matrix with 1 on the diagonal one can recover B_0 as follows: With $\Omega = CDC'$ and $\Omega = UU'$ its Cholesky decomposition, one finds that $C = U' \operatorname{diag}(\operatorname{diag}(U))^{-1}$ and

$$B_0 = -\left((U' \operatorname{diag}(\operatorname{diag}(U)))^{-1} - I\right).$$

The operator diag is employed to extract the diagonal elements of a square matrix as well as to construct a diagonal matrix from a vector.

2.2.4 Empirical Application

I present the approach detailed above for the initial example, namely, a vector autoregression in the spirit of Dimpfl & Jung (2012) in which the lag structure is determined as described in the previous section. To this end, one computes the realized volatility for all three markets for the in-sample period. In the next step, the adaptive lasso is employed as outlined above. For the adaptive weights ξ_i, I use the inverse of the magnitude of the OLS estimates, i.e., $\xi_i = |\beta_{i,OLS}|^{-1}$. This is a common choice and satisfies the conditions for the weights as required for the asymptotic result to hold.

Unlike in the standard univariate applications, where λ_T is determined by cross-validation, I determine λ_T using the Bayesian Information Criterion due to the lack of a reliable resampling scheme to justify cross-validation as is suggested by Kock & Callot (2012*b*). The maximal lag is set to $p = 10$ and $p = 20$ to assess the sensitivity of the procedure with regard to the maximal lag-length specification.

In summary, I estimate a model in the sense of (2.4) where β is a 90×1 (180×1) parameter vector. I employ the log of realized volatility which is commonly done in the realized volatility literature.[4] I then proceed to standardize the observations such that the magnitude of the coefficients can be interpreted. Recall that the purpose of this section is of pure in-sample (or exploratory) nature. At a later stage, I use the information gained at this stage to motivate the final model.

[4]For a justification beyond the usual log-normality argument of realized volatility in the case of linear models one can consult Audrino & Knaus (2012).

Table 2.1: Structural and reduced Form Estimates

Reduced Form			Structural Form		
			B_0	$\begin{pmatrix} . & . & . \\ 0.29 & . & . \\ 0.10 & 0.40 & . \end{pmatrix}$	$\begin{pmatrix} . & . & . \\ 0.28 & . & . \\ 0.10 & 0.41 & . \end{pmatrix}$
$\tilde{B}_1$	$\begin{pmatrix} 0.37 & 0.01 & 0.26 \\ . & 0.36 & 0.32 \\ . & . & 0.65 \end{pmatrix}$	$\begin{pmatrix} 0.39 & . & 0.25 \\ . & 0.38 & 0.33 \\ . & . & 0.68 \end{pmatrix}$	B_1	$\begin{pmatrix} 0.37 & 0.01 & 0.26 \\ -0.11 & 0.36 & 0.24 \\ -0.03 & -0.15 & 0.49 \end{pmatrix}$	$\begin{pmatrix} 0.39 & . & 0.25 \\ -0.11 & 0.38 & 0.26 \\ -0.04 & -0.15 & 0.52 \end{pmatrix}$
$\tilde{B}_2$	$\begin{pmatrix} 0.09 & . & . \\ . & 0.10 & . \\ . & . & 0.13 \end{pmatrix}$	$\begin{pmatrix} 0.10 & . & . \\ . & 0.08 & . \\ . & . & 0.13 \end{pmatrix}$	B_2	$\begin{pmatrix} 0.09 & . & . \\ -0.03 & 0.10 & . \\ -0.01 & -0.04 & 0.13 \end{pmatrix}$	$\begin{pmatrix} 0.10 & . & . \\ -0.03 & 0.08 & . \\ -0.01 & -0.03 & 0.13 \end{pmatrix}$
$\tilde{B}_3$	$\begin{pmatrix} 0.11 & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$	$\begin{pmatrix} 0.09 & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$	B_3	$\begin{pmatrix} 0.11 & . & . \\ -0.03 & . & . \\ -0.01 & . & . \end{pmatrix}$	$\begin{pmatrix} 0.09 & . & . \\ -0.02 & . & . \\ -0.01 & . & . \end{pmatrix}$
$\tilde{B}_4$	$\begin{pmatrix} . & . & . \\ . & 0.08 & . \\ . & . & 0.10 \end{pmatrix}$	$\begin{pmatrix} . & . & . \\ . & 0.08 & . \\ . & . & 0.10 \end{pmatrix}$	B_4	$\begin{pmatrix} . & . & . \\ . & 0.08 & . \\ . & -0.03 & 0.10 \end{pmatrix}$	$\begin{pmatrix} . & . & . \\ . & 0.08 & . \\ . & -0.03 & 0.10 \end{pmatrix}$
$\tilde{B}_6$	$\begin{pmatrix} . & . & -0.03 \\ . & . & . \\ . & . & . \end{pmatrix}$	$\begin{pmatrix} . & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$	B_6	$\begin{pmatrix} . & . & -0.03 \\ . & . & 0.01 \\ . & . & 0.00 \end{pmatrix}$	$\begin{pmatrix} . & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$
$\tilde{B}_7$	$\begin{pmatrix} 0.06 & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$	$\begin{pmatrix} 0.07 & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$	B_7	$\begin{pmatrix} 0.06 & . & . \\ -0.02 & . & . \\ -0.01 & . & . \end{pmatrix}$	$\begin{pmatrix} 0.07 & . & . \\ -0.02 & . & . \\ -0.01 & . & . \end{pmatrix}$
$\tilde{B}_8$	$\begin{pmatrix} 0.03 & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$	$\begin{pmatrix} 0.03 & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$	B_8	$\begin{pmatrix} 0.03 & . & . \\ -0.01 & . & . \\ 0.00 & . & . \end{pmatrix}$	$\begin{pmatrix} 0.03 & . & . \\ -0.01 & . & . \\ 0.00 & . & . \end{pmatrix}$
$\tilde{B}_9$	$\begin{pmatrix} . & . & . \\ . & . & . \\ 0.01 & . & . \end{pmatrix}$	$\begin{pmatrix} . & . & . \\ . & . & . \\ 0.01 & . & . \end{pmatrix}$	B_9	$\begin{pmatrix} . & . & . \\ . & . & . \\ 0.01 & . & . \end{pmatrix}$	$\begin{pmatrix} . & . & . \\ . & . & . \\ 0.01 & . & . \end{pmatrix}$
$\tilde{B}_{10}$	$\begin{pmatrix} . & . & -0.01 \\ . & 0.05 & . \\ . & . & 0.02 \end{pmatrix}$	$\begin{pmatrix} . & . & . \\ . & 0.03 & . \\ . & . & . \end{pmatrix}$	B_{10}	$\begin{pmatrix} . & . & -0.01 \\ . & 0.05 & 0.00 \\ . & -0.02 & 0.02 \end{pmatrix}$	$\begin{pmatrix} . & . & . \\ . & 0.03 & . \\ . & -0.01 & . \end{pmatrix}$
$\tilde{B}_{13}$	$\begin{pmatrix} . & . & . \\ . & . & . \\ . & 0.03 & . \end{pmatrix}$		B_{13}	$\begin{pmatrix} . & . & . \\ . & . & . \\ . & 0.03 & . \end{pmatrix}$	
$\tilde{B}_{15}$	$\begin{pmatrix} 0.05 & . & . \\ . & . & . \\ . & . & . \end{pmatrix}$		B_{15}	$\begin{pmatrix} 0.05 & . & . \\ -0.02 & . & . \\ -0.01 & . & . \end{pmatrix}$	

Estimates of B_j and $\tilde{B}_j$[a] for a maximal lag set to $p = 20$ (left column) and $p = 10$ (right column). Non-selected entries are marked by a dot. Indices j for which none of entries of either B_j or $\tilde{B}_j$ were selected are omitted.

[a] Note that for rounding reasons, $\tilde{B}_j$ of (2.2) cannot be reconstructed from B_0 and B_j.

The first important observation from Tab. 2.1 is that the estimates[5] do not change substantially when moving the maximal lag p from 10 to 20.

With this in mind, it is worthwhile looking at the magnitude of the coefficients at different lags: the reduced form parameters $\tilde{B}_i$ of Tab. 2.1 show a relatively rapid decline in size in which most of the off-diagonal entries are set to zero after the first lag. The structural parameters of Tab. 2.1 exhibit several remarkable facts. Since all the log $\mathrm{RV}_{i,t}$ are standardized, the (contemporaneous) impact of the FT on the SP, as measured by the size of the coefficient, is far larger than the Nikkei's impact on the S&P 500. This observation is in accordance with the actual provenance of the data, in the sense that the FT is temporally closer to the SP. The contemporaneous effect of the FT on the SP ($b_{0,32}$) is almost as large as the lagged value of the SP ($b_{1,33}$), while the lagged values both of the NE and of the FT are negligible when compared with the lagged values of the respective indices. Although the goal is to devise a model for the realized volatility of the SP, it is worth considering the rows different from the last one: the lagged influence of the SP on the FT ($b_{1,23}$) exceeds the influence of the NE on the FT ($b_{1,21}$), and the same holds true when reversing the roles of FT and NE. Additionally, the reduced parameters in $\tilde{B}_1$, clearly show the prominent role

[5]The omission of standard errors or p values is on purpose: Standard errors for-lasso based procedure are generally difficult if not impossible to obtain (Bühlmann & Van De Geer 2011) since the lasso provides biased estimates such the only viable way to obtain reasonable p-values would be – for the lack of an expression for the bias – to bootstrap. Since standard bootstrap schemes fail for the lasso (Chatterjee & Lahiri 2011) I refrain from providing standard errors or p values.

of the SP: elements of the third column (SP) is least often set to zero (for recent lags). This may be seen as evidence for the SP being the dominating market.[6]

Taken together, I consider this evidence to include the pseudo-closing realized volatility observation of the FT as an additional predictor for the realized volatility of the SP in the final model, in a sense to be specified more precisely in Section 2.4.1.

2.3 News Arrival

The impact of news on volatility has been studied from different angles: a pioneering work from the perspective of measuring news as perceived by returns is that of Engle & Ng (1993). The study of how returns affect news provides a deeper understanding of how returns and volatility are interlinked. This line of research, the "news impact curve" is extended to intraday data by Chen & Ghysels who find that "both very good news (unusual high intra-daily positive returns) and bad news (negative returns) increase volatility" (Chen & Ghysels 2011).

A different line of research addresses the impact of news in the sense of macroeconomic announcements and calendar day effects: this has already been documented for instance by Andersen, Bollerslev, Diebold & Vega (2007) for macroeconomic news, and calendar day effects for instance by Baillie & Bollerslev (1989). While most existing research looks at these effects from a GARCH-inspired perspective (using daily returns), Martens et al. (2009) extend this strand of research by suggesting a model, that features all of these effects in a realized volatility context. They devise a model for realized volatility that can accommodate leverage effects, macroeconomic announcements, weekly seasonality and level shifts in the volatility process.

In their contribution, Martens et al. (2009) assess different classes of models, ranging from a fractionally integrated AR process to an unrestricted AR model augmented with the

[6] As can be see in Tab. 2.1 some elements of the first column feature non-zero entries also at far legs. Given the magnitude of these coefficients and the lack of standard errors I consider these as noise.

aforementioned information. They conclude that despite the parsimonious and stringent specification of a fractionally integrated model, a flexible AR structure is superior. They also infer that the inclusion of this additional information is indeed beneficial except for level shifts.

Earlier works in this field find a high integration of stock markets around the globe by means of investigating the impact of U.S. macroeconomic news around the globe (Nikkinen, Omran, Sahlström & Äijö 2006). As a by-product they also find an effect of U.S. macroeconomic news on volatility using an extended GARCH model.

Building on Martens et al.'s (2009) results, I present two novel classes of models that feature the same augmented information set.

2.3.1 Existing Models

I consider Martens et al.'s (2009) specification as a yardstick with wich to compare the suggested classes. For this purpose, a variant of their model is revisited. To address the question of whether the incorporation of news, weekday and leverage effects is beneficial across model classes and whether this information can also be used in a simpler class of models, I present two extensions: the first extension is inspired by Corsi's (2009) HAR model, the second extension follows ideas presented in Audrino & Knaus (2012).

The first class of models (C_1), a conditional mean specification, is put forward by Martens et al. (2009) who postulate the following dynamics

$$y_t = \mu_t + \beta^{(d)}(y_{t-1} - \mu_{t-1}) + \beta^{(w)}(y_{t-1}^{(w)} - \mu_{t-1}^{(w)}) + \beta^{(m)}(y_{t-1}^{(m)} - \mu_{t-1}^{(m)}) + \epsilon_t \tag{2.5}$$

with $x_t^{(\kappa)}$ and $\mu_t^{(\kappa)}$ being aggregated frequencies as indicated by κ, which are either daily, weekly, or monthly. More precisely, $x_{t-1}^{(\kappa)} = \frac{1}{n_\kappa} \sum_{t=1}^{n_\kappa} x_{t-t}$ where n_κ is 1 (daily, $\kappa = d$), 5 (weekly, $\kappa = 5$), or 22 (monthly, $\kappa = m$). $\mu_{t-1}^{(\kappa)}$ is defined analogously. The contribution of

Martens et al. is not only the specification of a conditional mean model as such but also an elaborate specification of the process μ_t which is defined as follows:

$$\mu_t = c + \sum_{i=1}^{n_\omega} \omega_i \, w_{i,t} + \sum_{i=1}^{n_\eta} \eta_i \, n_{i,t} + \lambda r_{t-1}^{+} \tag{2.6}$$

$w_{i,t}$ is a dummy for the weekday of day t, $n_{i,t}$ is a dummy for a news arrival of type i on day t, and $r_{t-1}^{+} = \max(0, -r_{t-1})$ to account for the widely documented leverage effect.

Although Martens et al. (2009) also include fractionally integrated models and unrestricted models in their analysis, I second their conclusion that the benefit of modeling a fractionally integrated process does not produce a better description of the data. Also, an unrestricted AR(p) model included in their work performs no better than the parametrically imposed structure of (2.5) and (2.6).[7]

The second class (C_2), HAR-augmented models, is an extension of the model pioneered by Corsi (2009). Corsi's (2009) model suggests dynamics for the realized volatility, which is a weighted sum of the daily, weekly, and monthly volatility. These temporal components are augmented with the same exogenous factors as already found in the previous model, i.e.,

$$y_t = c + \beta^d y_{t-1}^{(d)} + \beta^w y_{t-1}^{(w)} + \beta^m y_{t-1}^{(m)} + \sum_{i=1}^{n_\omega} \omega_i \, w_{i,t} + \sum_{i=1}^{n_\eta} \eta_i \, n_{i,t} + \lambda r_{t-1}^{+} + \epsilon_t$$

where the variables are as defined for the first class of models. The coefficients are different across these models whereas the actual values of the additional regressors are identical.

The third class (C_3), an agnostic approach, is inspired by a different application of the lasso (as already found in Audrino & Knaus (2012)) applied to time series modeling. To this end, I employ the lasso procedure in the univariate time series case with a specific set

[7] In the methodological part of their work they also suggest a non-parametric estimate for c in (2.6), making it time dependent $c(t)$. The approach proposed is to model $c(t)$ non-parametrically, however, results for non-parametric $c(t)$ are not included in the paper which I attribute to the potentially poor identification of $c(t)$.

of predictors P that correspond to the predictors introduced for the two previous models, i.e.,

$$P = \{\{w_{i,t}\}_{1\leq i\leq n_\omega}, \{n_{i,t}\}_{1\leq i\leq n_\eta}, r^{+}_{t-1}, \{y_{t-i}\}_{1\leq i\leq 20}\}^{8}.$$

Unlike for the lasso as introduced in Section 2.2.3 or for the lasso as employed in Audrino & Knaus (2012), there is no theoretical framework that establishes model selection consistency in this context with exogenous regressors. Despite the lack of theoretical framework, this idea has already been successfully applied by Park & Sakaori (2013).

The selection of the shrinkage parameter uses a blocked cross-validation with a block length of 25 as suggested in Bergmeir & Benítez (2012).

To gauge the performance and appropriateness of the above specifications, I include an in-sample analysis that contrasts these three families in the same in-sample period. Again, this empirical application is meant to be exploratory and will later be used in Section 2.4.1.

2.3.2 In-sample Estimates

Unlike in the original work of Martens et al. (2009) I do not use centered dummies since the ultimate goal is to assess these models in terms of forecasting such that a reparametrization is irrelevant. Hence, one has for $i = 1, 2, 3, 4$

$$w_{i,t} = \begin{cases} 1 & \text{if weekday of date } t \text{ is day } i+1 \\ 0 & \text{otherwise} \end{cases}$$

where day 2 is Tuesday, 3 is Wednesday, 4 is Thursday, and 5 is Friday.

Following Martens et al. (2009) I use a dummy for macroeconomic news announcement for the GDP (Bureau of Economic Analysis, final GDP), unemployment figures (Bureau

[8] The lasso is employed in its usual cross-sectional variant to the equation $y_t = \sum_{x\in P} \beta_x x + \epsilon_t$.

of Labor Statistics, unemployment rate), inflation (Bureau of Labor Statistics, CPI), and federal fund rates (Federal Reserve).[9].

The in-sample estimates of these three models are collected in Tab. 2.2.

[9] In light of the non-standard actions taken by the FED in recent times, other announcements by the FOMC where included as additional predictors in an earlier version; these were not found significant and are thus excluded for the sake of consistency with earlier works in this version

Table 2.2: In-sample estimates of model classes 1 to 3

	C_1			C_2			C_3	Benchmark (HAR)		
	Estimate	Std. Error	p-Value	Estimate	Std. Error	p-Value	Estimate	Estimate	Std. Error	p-Value
Intercept	-9.774	0.820	0.000	-1.141	0.201	0.000	-1.397	-0.466	0.186	0.012
D	0.520	0.047	0.000	0.476	0.043	0.000	0.647	0.519	0.043	0.000
W	0.349	0.065	0.000	0.331	0.059	0.000	NA	0.344	0.060	0.000
M	0.085	0.036	0.018	0.099	0.032	0.002	NA	0.088	0.035	0.013
x^+_{t-1}	0.078	0.015	0.000	0.143	0.020	0.000	0.091			
Tuesday	0.155	0.037	0.000	0.230	0.051	0.000	0.062			
Wednesday	0.196	0.041	0.000	0.191	0.052	0.000	0.037			
Thursday	0.246	0.043	0.000	0.237	0.048	0.000	0.034			
Friday	0.147	0.055	0.008	0.115	0.070	0.101	-			
GDP	0.021	0.064	0.739	-0.036	0.060	0.545	-			
CPI	0.060	0.057	0.285	0.087	0.066	0.188	-			
FED	0.378	0.079	0.000	0.471	0.094	0.000	0.473			
EMP	0.183	0.067	0.006	0.188	0.068	0.006	0.053			
R^2			0.807			0.816	0.803			0.803
BIC			1'499.82			1'562.98	1'601.87			1'599.60
MSE			0.261			0.251	0.270			0.269
x_{t-2}							0.216			

Estimates of models class 1 to 3 as introduced in Section 2.4.1. Lags that are not selected by the lasso procedure are omitted. Note that the estimates for the weekly and monthly component are not available to the lasso. Moreover, x_{t-1} is the daily component. Standard errors are computed using a block-bootstrap scheme with 1000 replications and a block size of 25.

What can be observed from Tab. 2.2 is as follows: the autoregressive coefficients on the daily, weekly, and monthly components are of the expected magnitude, even for the conditional mean model. Moreover, the effect of weekdays is significant at the individual day level (as also reported by Martens et al. (2009)). However, the effect of news arrival is ambivalent: while higher volatility on federal fund rate and employment figure announcement days is common across models, the remaining news (inflation, GDP) are not found to be significant across both models. When it comes to intra-weekly seasonality, there is a slight difference across the two models: all weekday dummies are found to be significant in C_1, whereas the Friday-effect is not found to be significant in C_2.

These findings are also confirmed by C_3 in which the only predictors that are selected (beyond the leverage effect and the lagged observations) are the federal fund rate and employment figures and the weekday dummies (excluding Friday).

These results are in line with what is reported in the online appendix of Martens et al. (2009) for fractionally integrated models, with the exception that they find that the release of inflation data as measured by the CPI dummy is significant as well. However, I attribute this difference to either the sample or – more importantly – to the different nature of the models.

With these in-sample estimates at hand I reach the following conclusions: there is a clear indication that a model featuring leverage effects and intra-week seasonality is worthwhile considering and the performance difference between C_1 and C_2 is not substantial from an in-sample point of view. The role of macroeconomic news appears relevant only for employment and federal fund rate announcements. Given the results found herein combined with the conclusion of Martens et al. (2009) these results will be used to construct the final set of models found in 2.4.1.

2.4 Augmented Linear Models

In view of what has been addressed in Section 2.2 and Section 2.3 of this paper I now consider the question if and to what extent volatility spillovers and news (in the wide sense) are beneficial to volatility forecasting. Reviewing the in-sample results and the models introduced I propose three classes of models and assess their performance in an out-of-sample comparison.

I impose further restrictions to understand the benefit of each component with regard to volatility forecasting. The choice and specification of the models is guided by the knowledge gained in the preceding sections of the current paper. All the information collected in this in-sample period is taken to the forecasting comparison outlined in the next section.

2.4.1 Competing Models

I use the models of Class 1 to 3 introduced in Section 2.3 but augment Class 2 and Class 3 by the volatility of other markets as introduced in Section 2.2. Since the conditional mean specification does not allow for a natural inclusion of lagged volatilities of other markets I refrain from doing so in Class 1.

In summary, I include the following models in the out-of-sample prediction exercise.

Class 1: As specified in (2.5) with μ_t

$$\mu_t = c + \delta_1 \sum_{i=1}^{n_\omega} \omega_i \, w_{i,t} + \delta_2 \sum_{i=1}^{n_\eta} \eta_i \, n_{i,t} + \delta_3 \lambda r_{t-1}^{+}. \tag{2.7}$$

δ_k takes values in $\{0, 1\}$ for $k = 1, \ldots, 3$. This ensures that only a specific group of additional information is included in the model. However, I impose that $\sum_{k=1}^{3} \delta_k > 0$: otherwise, for $\delta_k = 0$ for all $k = 1, 2, 3$, the model collapses to the usual HAR model. Hence, there is a total of 7 models for Class 1.

Class 2:

$$\begin{aligned} y_t = c + \beta^d y_{t-1}^{(d)} + \beta^w y_{t-1}^{(w)} + \beta^m y_{t-1}^{(m)} + \delta_1 \sum_{i=1}^{n_\omega} \omega_i \, w_{i,t} \\ + \delta_2 \sum_{i=1}^{n_\eta} \eta_i \, n_{i,t} + \delta_3 \lambda r_{t-1}^{+} + \delta_4 \sum_{i=1}^{l} \phi_i y_{t-h_i}^{i,(a_i)} + \epsilon_t \end{aligned} \tag{2.8}$$

$y_{t-h_i}^{i,(a_i)}$ is the volatility of market m_i, lagged by h_i and aggregated into a_i. In light of the results of Section 2.2 the sum is limited to $l = 1$ with $m_1 = 2$ (the FT), $h_1 = 0$ and $a_1 = 1$ such that (2.8) only includes the realized volatility of the FT of the current day. Similar to class 1, δ_k takes values in $\{0, 1\}$ for $k = 1, \ldots, 4$ such that Class 2 yields a total of 16 models.

Class 3: Again the lasso procedure is employed with

$$\{\{w_{i,t}\}_{1 \le i \le 4}, \{n_{i,t}\}_{1 \le i \le 5}, r_{t-1}^{+}, \{y_{t-i}\}_{1 \le i \le 10}, y_t^{2,(1)}\}$$

as available predictors, i.e., the weekdays, macroeconomic news, leverage indicator, as well as lagged volatilities of the last 10 days.

Models 1 to 3 are summarized by $C_{1,\delta_1\delta_2\delta_3\delta_4}$, $C_{2,\delta_1\delta_2\delta_3}$ and C_3. Lastly, I also include what has been laid out in Section 2.2 as a class of its own, Class 4. The vector autoregression by means of the lasso introduced in Section 2.2.3 can straightforwardly be extended to out-of-sample forecasting and is denoted as C_4.

Macroeconomic news is for all the models limited to announcements of the employment rate and the federal fund rate; in line with the results of Section 2.3. The same argument applies to the maximal lag of past volatilities for C_3 and C_4 which is set to $p = 10$ as well as to only including the FT as a predictor (as found in Section 2.2 and 2.3).

2.4.2 Out-of-sample Evaluation

Compliant with the ultimate goal of daily volatility *forecasting*, I present a volatility forecasting study in the two subsequent paragraphs. To maintain strictly the idea of training and test data (Hastie et al. 2009), none of the data on which the models are evaluated, have been used in the preceding sections. Hence, the results presented hereafter are realistic estimates of the performance of these models in a real-world application.

The out-of-sample evaluation starts on January 3, 2011 and ends September 25, 2012, resulting in 437 out-of-sample evaluation points. For a more detailed description the reader is again referred to the appendix.

The volatility forecasting evaluation looks at two aspects: first a comparison in the sense of simple out-of-sample forecasting is presented, and, second, a more practical value-at-risk forecasting evaluation is conducted.

Forecasting Comparison

Since there is no prior assumption on which model will fare best, I resort to the model confidence set approach put forward by Hansen, Lunde & Nason (2011). The model confidence set (MCS) – to a certain extent similar to a confidence interval for a parameter estimate – contains the best model with a certain probability $1 - \alpha$: the estimated model confidence set at level $1 - \alpha$, $\widehat{\mathcal{M}}^*_{1-\alpha}$ contains the true best models with a probability of at least $1 - \alpha$. Hence, the smaller α the more elements are contained in the final model confidence set. An important feature of the MCS approach is that it accounts for the source of the data and potentially delivers a large set if the data is uninformative. More precisely, suppose that there is loss function L that assigns a loss to the prediction $\hat{y}_{t,i}$ of a model i with regard to the realization y_t, i.e., $L_{i,t} = L(y_t, \hat{y}_{i,t})$. Define $\mu_{ij} = \mathbb{E}(L_{i,t} - L_{j,t})$ as the expected relative performance of model i over model j. The set of superior models is defined as

$$\mathcal{M}^* = \{i \in \mathcal{M}^0 : \mu_{ij} \leq 0 \; \forall j \in \mathcal{M}^0\}$$

where $\mathcal{M}^0$ is the initial set of models. To obtain an estimate of $\widehat{\mathcal{M}}^*_{1-\alpha}$, i.e., the set that contains the true superior set with probability $1 - \alpha$ while controlling the nominal level in this multiple comparison setting, Hansen et al. suggest (among others) an approach based on the procedure of testing for equal predictive ability $\mathcal{H}_0 : \mu_{ij} = 0 \quad \forall i, j \in \mathcal{M}$. The procedure starts off with $\mathcal{M} = \mathcal{M}^0$ and iteratively eliminates a model from $\mathcal{M}$; this procedure is reiterated until $\mu_{ij} = 0$ cannot be rejected. The test and elimination rules are based on

$$T_{\max,\mathcal{M}} = \max_{i \in \mathcal{M}} t_{i\cdot} \quad \text{with } t_{i\cdot} = \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{\text{Var}}(\bar{d}_{i\cdot})}}$$

where $\bar{d}_{i\cdot} = \frac{1}{\#\mathcal{M}} \sum_{j \in \mathcal{M}} \frac{1}{n} \sum_{t=1}^{n} L_{i,t} - L_{j,t}$. To find the distribution of $T_{\max,\mathcal{M}}$ Hansen et al. suggest a bootstrap procedure based on the block bootstrap. I follow this suggestion and use the block bootstrap with a block length of 25 and 20'000 replications. Concerning the loss function, the analysis is primarily based on the squared predictive error, i.e. $L(y_t, \hat{y}_{i,t}) = (y_t - \hat{y}_{i,t})^2$ as well as the absolute predictive error $L(y_t, \hat{y}_{i,t}) = |y_t - \hat{y}_{i,t}|$ such that the mean squared predictive error (MSPE) and the mean absolute predictive error (MAPE) can be reported for the whole evaluation sample. [10] Tab. 2.3 thus reports the MSPE and MAPE together with the corresponding MCS p-values as well as the p-values of the Mincer-Zarnowitz regression of the null hypothesis of unbiasedness (Mincer & Zarnowitz 1969) of the forecast (both encoded with stars).

The first remark concerns the robustness of the results presented in Tab. 2.3. The model confidence sets agree to a great extent for the two loss functions, moreover, the results do not change dramatically when replacing the estimate of RV_t with the naive estimator as shown in Tab. B.4. In addition, the sample size is of minor importance for the relative

[10] There is research identifying loss functions (Patton 2011) that are robust if the true realization of volatility is observed with errors. This cannot be applied in the present case since the modeled quantity is the log volatility, consequently, I restrict the analysis to the two loss function introduced earlier.

Table 2.3: Out-of-sample results

	δ_1	δ_2	δ_3	δ_4	MSPE	Var	Bias	MAPE	R^2
C_1	1	0	0	-	0.281	0.281	−0.016%	0.422	0.670
	0	1	0	-	0.285	0.285	−0.018%	0.423	0.665
	1	1	0	-	0.274	0.274	−0.014%	0.416	0.678
	0	0	1	-	0.278	0.278	−0.017%	0.419	0.673
	1	0	1	-	0.270	0.270	−0.013%	0.416	0.682
	0	1	1	-	0.275	0.275	−0.015%	0.416	0.676
	1	1	1	-	0.264	0.265	−0.011%	0.410	0.688
C_2	0	0	0	0	0.290	0.291	−0.052%	0.426	0.658
	1	0	0	0	0.281	0.282	−0.024%	0.423	0.669
	0	1	0	0	0.264	0.264	0.020%	0.411	0.689
	1	1	0	0	0.231***	0.229	0.343%	0.381***	0.732**
	0	0	1	0	0.273	0.273	−0.004%	0.416	0.678
	1	0	1	0	0.239**	0.237	0.332%	0.386**	0.722**
	0	1	1	0	0.281	0.281	−0.043%	0.419	0.670
	1	1	1	0	0.255	0.255	0.023%	0.403	0.699
	0	0	0	1	0.247	0.245	0.343%	0.392	0.714**
	1	0	0	1	0.239**	0.236	0.353%	0.388*	0.724***
	0	1	0	1	0.232***	0.230	0.348%	0.379***	0.732***
	1	1	0	1	0.265	0.265	0.003%	0.409	0.688
	0	0	1	1	0.239**	0.237	0.343%	0.383***	0.724***
	1	0	1	1	0.271	0.271	−0.018%	0.414	0.682
	0	1	1	1	0.232***	0.230	0.334%	0.378***	0.731**
	1	1	1	1	0.225***	0.223	0.339%	0.373***	0.739**
C_3	-	-	-	-	0.232***	0.229	0.391%	0.380***	0.732***
C_4	-	-	-	-	0.248	0.247	0.239%	0.398	0.710**

The stars of MSPE and MAPE are MCS p-values. Models with one star are in $\widehat{\mathcal{M}}^*_{0.95}$, with two stars in $\widehat{\mathcal{M}}^*_{0.9}$, with three stars in $\widehat{\mathcal{M}}^*_{0.75}$. Note that $\widehat{\mathcal{M}}^*_{0.75} \subset \widehat{\mathcal{M}}^*_{0.9} \subset \widehat{\mathcal{M}}^*_{0.95}$. The stars superscripted with R^2 are the p-value of the Mincer-Zarnowitz regression, where $***$ corresponds to 0.01, $**$ to 0.05, and $*$ and 0.1.

results: while the model confidence sets obtained in Tab. B.3 are larger for smaller window sizes (due to the lack of power of the MCS procedure induced by the high variance of the forecast error), the relative results are stable across training window sizes. Looking at the absolute value of the MSPE (MAPE) across training window sizes, I consider a training window size of 1000 observations to be a reasonable size, which is also often found in comparable studies.

With this in mind, a closer look at the results is warranted. When examining the models in the smallest model confidence set, $\widehat{\mathcal{M}}^*_{0.75}$, the model with the lowest MSPE (MAPE) is indeed the saturated model of class C_2, i.e., $C_{2,1111}$, however, the model $C_{2,1100}$ is a close second hinting (only MSPE) at the possibility that weekly seasonality and macro announcements are sufficient to attain reasonably low MSPEs, however, the models $C_{2,0101}$, $C_{2,0111}$ and C_3 are also found to be in this set. The interpretation of the remaining results is even less trivial: when considering models that are only augmented by one group of predictors (i.e., $\delta_k = 0$ except for one k) the model featuring the FT as predictor ($C_{2,0001}$) performs best. A next interesting observation is that $C_{2,1001}$ and $C_{2,0101}$, models with weekday and macro news augmented with the FT, deliver good results. These observations taken together may indicate that the role of the FT is that of a panacea, or, bluntly speaking, a "catch-all" variable (also in the sense that it may capture the leverage effect through past returns of the SP affecting the FT and thus again the SP). The same argument applies to $C_{2,0011}$ in which the potential role of intra-week seasonality and macro news effects is taken over by the role of the FT (δ_4). It can thus be argued that only $C_{2,1111}$ (or alternatively C_3) are able to capture the net effect of the inclusion of the FT.

A further observation, worth highlighting is the result for C_3 and C_4: although they do not feature the lowest MSPE, they fare considerably well. Most importantly, they both outperform the basic HAR model ($C_{2,0000}$) significantly with t-statisics of -4.364 (C_3) and -3.235 (C_4) in a test of equal predictive power (Diebold & Mariano 1995) with HAC robust standard errors (Newey & West 1987, Newey & West 1994). This is particularly noteworthy since both have a maximal lag of $p = 10$ whereas the HAR model has an implicit maximal lag of $p = 22$ to accommodate the apparent near long-memory features of the time series.

The last remark concerns the absolute value of the coefficient-of-determination: although comparing the R^2 across different studies (using different data, sample sizes, etc.) cannot be considered a reliable benchmark, I still deem it worthwile mentioning that the highest R^2 reported by Chen et al. (2010), comparing several models at the one day horizon for log RV, is 0.718, while the HAR model's R^2 in the aforementioned study is found to be 0.691, thus hinting at the possibility that a saturated model ($R^2 = 0.739$) is indeed beneficial to forecasting daily realized volatility, even when compared with larger set of competitors.

When examining the MSPE in terms of bias and variance, it becomes apparent that all the models of class C_1 have a considerably lower bias as opposed to models in C_2, C_3, and C_4: this is not unexpected given the specification of the model in terms of a *conditional mean* which comes at the price of higher variance of the forecast error. Also, all the models in $\widehat{\mathcal{M}}^*_{0.75}$ have a considerably lower variance which comes at the price of higher bias.

Looking at the bias also leads to the most perplexing result of this out-of-sample comparison. While the results in the model confidence set approach appear robust across different loss functions (MSPE, MAPE), the test for unbiasedness is often rejected for models included in the model confidence set. Although, this seems perplexing at first glance it can be explained by the null hypotheses of the respective procedure: The model confidence set delivers a set of best models (in relative terms to competing models), Mincer & Zarnowitz's (1969) approach tests for unbiasedness (in absolute terms, ignoring competing models). Thus, the relatively high variance and low bias for models of class C_2 are not rejected in terms of unbiasedness, whereas other models get rejected in Mincer & Zarnowitz's (1969) test for reasons of high bias and low variance.

In summary, it can be stated that the joint inclusion of supplemental information lowers the MSPE regardless of the class of models, additionally, much shorter lags (as found in C_3 and C_4) are sufficient to attain reasonably low prediction errors. Another relevant point at this stage is that only with the inclusion of macroeconomic news and intra-weekly seasonality one already has a lower MSPE highlighting the role of exogenous effects in volatility models.

The question of whether bias or variance is the more important constituent of the MSPE depends of course on the application in question. To answer this question in part I consider a value-at-risk application in the next section.

Value-at-risk Application

To assess the performance of the volatility forecasts from a more practical perspective I compute strict out-of-sample value-at-risk estimates for the same out-of-sample period for different levels of the value-at-risk. As in the preceding paragraph, I limit the forecasts to daily value-at-risk forecasts, first, because of the reasons mentioned earlier, and second, more importantly, because of the currently prevailing value-at-risk regulations that require a daily value-at-risk computation.[11] The one-period ahead value-at-risk at the level α is defined as

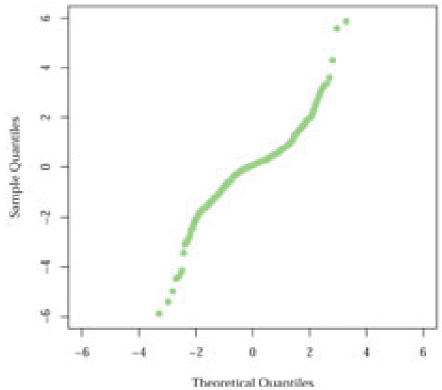
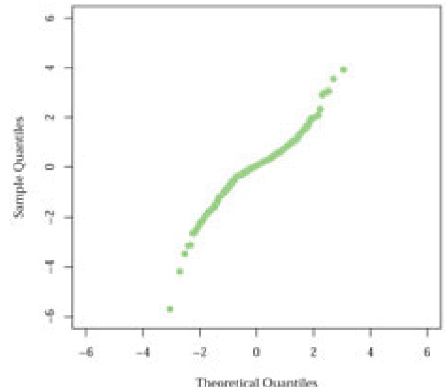
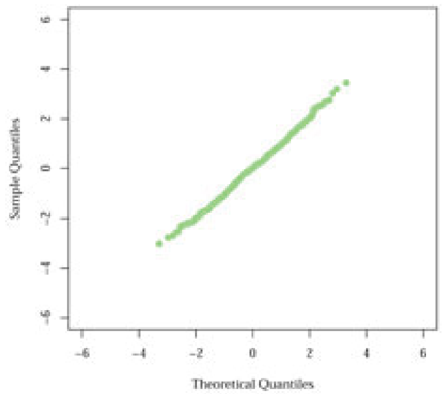
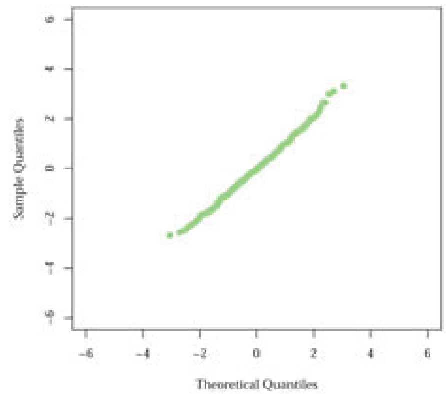
$$\mathrm{VaR}^{\alpha}_{t+1} = \inf_{\xi}\{\xi \in \mathbb{R} \mid \mathrm{P}(r_{t+1} \leq \xi | \mathcal{F}_t) \geq \alpha\}. \tag{2.9}$$

In this case $\mathcal{F}_t$ is the σ-algebra containing all the information up to date t (including the partial day volatility of the FT). Actually, to predict $\mathrm{P}(r_{t+1} \leq \xi | \mathcal{F}_t)$ one needs an assumption on the distribution of r_{t+1} conditional on $\mathcal{F}_t$. In the present case the assumption that $r_t / \sqrt{\mathrm{RV}_t}$ is normally distributed is employed. The results presented in Fig. 2.2 show no evidence of this assumption being violated in the data at hand. [12]

[11] Most recent literature compares daily value-at-risk forecast (Halbleib & Pohlmeier 2012). To extend the coverage period of the value-at-risk, the daily forecast is scaled with the square root of the time horizon.

[12] For the sake of in- and out-of-sample analysis I report these separately in Fig. 2.2 for both sample periods along with p values of of the Shapiro-Wilk and Jarque-Bera test for normality (Shapiro & Wilk 1965, Jarque & Bera 1980).

Figure 2.2: Quantile-quantile plot of (standardized) scaled logarithmic returns

	In-sample	Out-of-sample
r_t		
$\frac{r_t}{\sqrt{\mathrm{RV}_t}}$		
Shapiro-Wilk	0.241	0.541
Jarque-Bera	0.409	0.253

Quantile-quantile plot of (standardized) scaled logarithmic returns against normal distribution with p values of Normality tests.

As can be seen from Fig. 2.2 the raw log returns (scaled) feature the usual skewness and excess kurtosis, while the log returns standardized by the realized volatility (scaled) look fairly Gaussian; this is confirmed by the p-values of the normality tests.[13] Since there is no *ex-ante* reason to reject the assumption of normality I maintain this assumption and proceed to compute the value-at-risk as outlined in (2.9).[14]

A further problem that becomes apparent when attempting to forecast value-at-risk with realized volatility is the fact that realized volatility is modeled as its logarithm. While any model introduced in 2.4.1 produces a forecast for $\mathbb{E}_t(y_{t+1})$ (with y_{t+1} being log realized volatility), a forecast of the transformed series $\exp(y_{t+1})$ is required. Since the logarithm is a special case of the Box-Cox transform (Box & Cox 1964) one can draw on the research of Proietti & Lütkepohl (2013) (and the earlier work of Guerrero (1993) and Pankratz & Dudley (1987)) who establish that the bias induced by the transformation can be accounted for if one is willing to make the assumption that the transformed series is normally distributed. Despite the fact that a formal test rejects the normality assumption, this assumption is still maintained: first, the deviation from normality is not too pronounced as shown in Fig. B.1, and, second, this allows for an explicit computation of the bias

[13] This finding is stable across different assets, for instance for foreign exchange returns (Andersen, Bollerslev, Diebold & Labys 2000) and returns of the FTSE 100 futures (Areal & Taylor 2002).

[14] While this result is not surprising if one assumes the classical diffusion for the asset price, it is in fact surprising to hold for the bipower variation. To reconcile this result with existing research, I hypothesize that while jumps are present at the intra-day level, their impact on daily log-returns appears negligible (Barndorff-Nielsen & Shephard 2004).

correction.[15] The conditional expectation (which minimizes the MSPE by construction) of $\exp y_{t+1}$ can then be computed as follows

$$\mathbb{E}_t(\exp(y_{t+1})) = \widehat{\exp y_{t+1}} = \exp(\widehat{y_{t+1}})\exp(\sigma^2_{t+1}/2).$$

where σ^2_{t+1} is the variance of the forecast error.

Hence, one can compute the value-at-risk based on a forecast of model M as

$$\text{VaR}^{1-\alpha}_{M,t+1} = \Phi^{-1}(1-\alpha)\sqrt{\exp(\widehat{y_{M,t+1}})\exp(\sigma^2_{M,t+1}/2))} + \mu_t \tag{2.10}$$

where $\widehat{y_{M,t+1}}$ is the forecast of the log realized volatility produced by model M, $\sigma^2_{M,t+1}$ its associated forecast error variance, and $\mu_t = \mathbb{E}_t(r_{t+1})$ the conditional expectation of the returns. Both $\sigma^2_{M,t+1}$ and μ_t are assumed to be constant over time such that these can be estimated over the training window.[16] Φ^{-1} is the quantile function of a standard normal distribution.

To evaluate the value-at-risk forecasts produced by the different models, the number of violations (hit ratio, HR) of a specific model M is computed, i.e.

$$\text{HR}^{1-\alpha}_M = \#\{r_{t+1} < \text{VaR}^{1-\alpha}_{M,t+1}\}/n$$

where n is the total number of out-of-sample evaluations. If the model is well specified one should obviously find $\text{HR}^{1-\alpha}_M \simeq 1-\alpha$. In accordance with the existing literature (Kuester, Mittnik & Paolella 2006, Martens et al. 2009, Halbleib & Pohlmeier 2012) I employ a test of correct unconditional coverage (Kupiec 1995) as well as an extension of Kupiec's (1995) test to assess the conditional coverage probability developed by Christoffersen (1998). The former simply tests whether the unconditional hit ratio is violated with a likelihood ratio test, whereas the latter tests the joint hypothesis of independence (against a first-order

[15] As already mentioned earlier, the use of logarithm to transform realized volatility is arbitrary. A more thorough approach may be to use the Box-Cox transform and determine the transformation parameter in data-driven way (see for instance Proietti & Lütkepohl (2013). Preliminary results show that for $\lambda \simeq -0.2$ in the Box-Cox transform, Normality of the unconditional distribution cannot be rejected. In order to make these results comparable to existing research I nonetheless adopt the log transform, while I am inclined to believe that the use of a proper power transform to model realized volatility bears great potential (see for istance Gonçalves & Meddahi (2011))

[16] This approach has already been followed by Giot & Laurent (2004), however, without the context of the Box-Cox transform. The assumption of constant forecast error variance is already found in Giot & Laurent (2004) and Proietti & Lütkepohl (2013)

Markovian dependence) and correct unconditional coverage, yielding a likelihood ratio test for the correct conditional coverage (LR_{UC}). In more detail,

$$\mathrm{LR}_{UC} = 2 \cdot (\log \mathcal{L}_1 - \log \mathcal{L}_0)$$

where $\mathcal{L}_0 = \alpha^k(1-\alpha)^{n-k}$, $\mathcal{L}_1 = \hat{\alpha}^k(1-\hat{\alpha})^{n-k}$ with $k = \#\{r_{t+1} < \mathrm{VaR}^{1-\alpha}_{M,t+1}\}$ and $\hat{\alpha} = k/n$. LR_{UC} can be shown to follow a χ_1-distribution (Kupiec 1995). Testing against first order Markov dependence can be performed in the following way with

$$\mathrm{LR}_{IND} = 2 \cdot (\log \mathcal{L}_D - \log \mathcal{L}_0)$$

where $\mathcal{L}_0$ is as defined above and $\mathcal{L}_D = (1-\pi_{ne,e})^{k_{ne,ne}}\pi_{ne,e}^{k_{ne,e}}(1-\pi_{e,e})^{k_{e,ne}}\pi_{e,e}^{k_{e,e}}$ where $k_{\cdot,\cdot}$ is the number of observations where an exceedance (non-exceedance) is followed by an exceedance (non-exceedance) with exceedance codified as *e* and non-exceedance codified as *ne*. The relevant transition probabilities are straightforwardly estimated as $\pi_{ne,e} = \frac{k^{ne,e}}{k^{ne,e}+k^{ne,ne}}$ and $\pi_{e,e} = \frac{k^{e,e}}{k^{e,ne}+k^{e,e}}$. LR_{IND}, the likelihood ratio test against dependence, is again distributed as χ_1. Lastly, when conditioning on the first observation one finds that the likelihood ratio test of the joint hypothesis, the correct conditional coverage (LR_{CC}), is given as $\mathrm{LR}_{CC} = \mathrm{LR}_{UC} + \mathrm{LR}_{IND}$ which is χ_2 distributed (Christoffersen 1998).

The extension of Christoffersen is crucial: while the unconditional correct coverage is interesting, a correctly specified model should indeed yield value-at-risk forecasts that are correct and independent, thus, the conditional correct coverage should examined. With these tests at hand I now turn to the out-of-sample evaluation of the value-at-risk forecasts.

The out-of-sample results for all the models introduced in 2.4.1 are collected in Tab. 2.4 for $\alpha = 1\%, 2.5\%$.[17] Along with the hit ratios, I report the *p*-values of the correct unconditional coverage (UC), independence (IND), and correct conditional coverage (CC) test in Tab. 2.4. To contrast the estimates produced using the (log) realized volatility models I include a naive *static* forecast which simply uses the empirical $1-\alpha$ quantile over the training window width.

[17] An earlier version also included $\alpha = 5\%, 10\%$. The conclusions drawn in that setting did not differ from the conclusions drawn with the current nominal levels of $\alpha = 1\%, 2.5\%$ which I deem more appropriate from a risk management perspective.

Table 2.4: Out-of-sample value-at-risk forecasts

					VaR 1%				VaR 2.5%			
	δ_1	δ_2	δ_3	δ_4	Hit Ratio	UC	IND	CC	Hit Ratio	UC	IND	CC
C_1	1	0	0	-	3.43%	0.00	**0.30**	0.00	4.81%	0.01	**0.15**	0.01
	0	1	0	-	3.43%	0.00	**0.30**	0.00	4.81%	0.01	**0.15**	0.01
	1	1	0	-	3.43%	0.00	**0.30**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	0	0	1	-	3.66%	0.00	**0.27**	0.00	4.58%	*0.01*	**0.17**	*0.02*
	1	0	1	-	3.66%	0.00	**0.27**	0.00	4.58%	*0.01*	**0.17**	*0.02*
	0	1	1	-	3.43%	0.00	**0.30**	0.00	4.81%	0.01	**0.15**	0.01
	1	1	1	-	3.43%	0.00	**0.30**	0.00	4.35%	*0.02*	**0.19**	*0.03*
C_2	0	0	0	0	3.66%	0.00	**0.27**	0.00	4.81%	0.01	**0.15**	0.01
	1	0	0	0	3.66%	0.00	**0.27**	0.00	4.81%	0.01	**0.15**	0.01
	0	1	0	0	3.66%	0.00	**0.27**	0.00	4.58%	*0.01*	**0.17**	*0.02*
	1	1	0	0	3.43%	0.00	**0.30**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	0	0	1	0	3.66%	0.00	**0.27**	0.00	4.58%	*0.01*	**0.17**	*0.02*
	1	0	1	0	3.43%	0.00	**0.30**	0.00	4.58%	*0.01*	**0.17**	*0.02*
	0	1	1	0	3.43%	0.00	**0.30**	0.00	4.81%	0.01	**0.15**	0.01
	1	1	1	0	3.43%	0.00	**0.30**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	0	0	0	1	3.43%	0.00	**0.30**	0.00	4.58%	*0.01*	**0.17**	*0.02*
	1	0	0	1	3.43%	0.00	**0.30**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	0	1	0	1	3.20%	0.00	**0.34**	0.00	4.58%	*0.01*	**0.17**	*0.02*
	1	1	0	1	3.43%	0.00	**0.30**	0.00	4.81%	0.01	**0.15**	0.01
	0	0	1	1	3.43%	0.00	**0.30**	0.00	4.58%	*0.01*	**0.17**	*0.02*
	1	0	1	1	3.43%	0.00	**0.30**	0.00	4.81%	0.01	**0.15**	0.01
	0	1	1	1	3.43%	0.00	**0.30**	0.00	4.58%	*0.01*	**0.17**	*0.02*
	1	1	1	1	3.20%	0.00	**0.34**	0.00	4.35%	*0.02*	**0.19**	*0.03*
C_3	-	-	-	-	2.97%	0.00	**0.37**	0.00	4.35%	*0.02*	**0.19**	*0.03*
C_4	-	-	-	-	3.43%	0.00	**0.30**	0.00	4.12%	*0.05*	**0.21**	**0.06**
Static	-	-	-	-	0.23%	**0.05**	**0.92**	**0.15**	0.69%	0.00	**0.81**	*0.02*

The hit ratio (HR) together with the p-values of tests of correct unconditional coverage probability (UC), independence (IND), and conditional coverage probability (CC) are shown. Non-rejection at the 0.1 (0.05) level is highlighted in boldface (italics). Models are encoded as introduced in 2.4.1, the last column is the naive value-at-risk estimator using the empirical quantiles.

When drawing conclusions from these results it is important also to consider the results reported in Tab. B.6 which are equivalent to Tab. 2.4 with the difference that the entire analysis has been carried out using the naive estimator instead of the bipower variation.[18]

The conclusions that can be drawn from the out-of-sample value-at-risk predictions reported in Tab. 2.4 and Tab. B.6 are not evident: clearly, using the naive estimator (Tab. B.6) one finds generally more accurate value-at-risk forecasts and hit ratios that are closer to

[18] This is crucial as the value-at-risk quintessentially depends on $P(r_{t+1} \leq \alpha | \mathcal{F}_t)$, i.e., the distribution of r_t. Although the deviation of $r_t / \sqrt{RV_t}$ from Gaussianity is not rejected (as shown in Fig. 2.2), the deviation from the straight line in a quantile-quantile plot when standardizing with the naive estimator for RV_t is even less pronounced (see Fig. B.4)

the nominal level, thus resulting in higher p-values. Reassuring across all the models and levels is the non-rejection of the independence assumption: there appears to be no dependence in the value-at-risk violations.

A comparison among the models is even less evident since the comparison hinges on a few observations. Nonetheless, it can be ascertained from Tab. B.6, that the models for VaR 1% featuring the highest p-values in the correct conditional coverage test all contain FT as an additional predictor. When looking at the models that have an empirical hit ratio closest to the nominal level of 2.5%, two of these ($C_{2,1001}$, C_4) also contain FT as predictor, however, also $C_{1,001}$ achieves the same level.

These observations warrant some further inspection: first, the observation that most hit ratios using the naive estimator are closer to the nominal hit ratio, is likely to be attributed to the fact that returns, standardized by the naive estimator instead of the bipower variation, deviate less from a straight line in a quantile-quantile plot (see Fig. 2.2 and Fig. B.4). Second, the benefit of including the FT can potentially be explained by looking at the distribution of returns standardized by the forecasted volatility. Fig. B.5 collects these quantile-quantile plots for a selection of models (again using the naive estimator). What can be observed from Fig. B.5 is a kink at the lower left end which corresponds to volatility that is estimated too low on days with low returns. This kink can partly be "straightened" when including the FT as a predictor. Additionally, $C_{2,1100}$, which fared well in the pure forecasting exercise, displays returns that are much closer to the normal distribution than for instance $C_{2,0000}$. This again hints at the possibility that the FT takes the role of a panacea, which may indeed be beneficial when forecasting volatility on low return days.

Hence, it appears worthwhile to include additional predictors if the ultimate goal is value-at-risk forecasting. Put differently, the non-normality (partly induced by the under-estimation of volatility on low return days) of returns standardized by forecasted volatility can be reduced.

While the relative performance of the value-at-risk forecasts compared with each other cannot be assessed in a statistically rigorous way[19] it can be retained that, including the FT as an additional predictor may be beneficial (particularly if compared with the standard HAR model, $C_{2,0000}$) to value-at-risk forecasts. This again fits well (King & Wadhwani 1990) with the observation that negative news tend to affect markets jointly more severely.

The answer to the question raised in Kuester et al. (2006) of whether value-at-risk forecasts based on realized volatility deliver any benefit over models using daily data is beyond the scope of this paper. In light of the results above which suggest that there is an indeed an improvement in terms of value-at-risk when using augmented volatility models this may need to be reinvestigated. Further refinements of this procedure (e.g., using filtered historical simulation to approximate the return density) are left for future research.

2.5 Discussion and Conclusion

Revisiting a simple and parsimonious VAR model with the lasso leads to the conclusion that the inclusion of foreign markets is beneficial for volatility forecasting. An inspection of the lags of this lasso VAR regression adds evidence for the SP being the dominant market. Further, I have provided evidence in this paper that the inclusion of macroeconomic news formalized as a dummy for the announcement days has a robust impact across different models; these are also found to be relevant from a model selection perspective using the lasso. Taking these two findings together in a model that features both news (macroeconomic announcement, leverage effects, intra-week seasonality) and spillovers I find a superior out-of-sample performance of this model. The role of spillover is however less clear in this combined setting. Given the potentially dominant role of the S&P 500 index over other indices, the inclusion of the FTSE 100 may well be a panacea for effects (reaction to news, leverage effects, seasonality) present in the S&P time series. Nonetheless, a saturated model performs best in an out-of-sample setting and evidence of good performance of models with a considerably shorter lag structure than the HAR model is witnessed.

[19] A procedure like the model confidence set produces uninformative results due to the lack of power induced by the small value of $\alpha \cdot n$.

The question of whether it is sufficient to model realized volatility in an autarkic manner (Chen et al. 2010, Asai, McAleer & Medeiros 2012, Corsi & Renò 2012, Audrino & Knaus 2012) or whether the inclusion of additional predictors is required (Martens et al. 2009, Scharth & Medeiros 2009) cannot be ultimately answered in this paper. Nonetheless, I find strong evidence that within the classes revisited in the present study, the inclusion indeed appears beneficial. The last question, only tangentially relevant to this paper, which appears intriguing, is the question of power transforming realized volatility series and its impact on forecasting and derived applications (e.g., value-at-risk). Valuable future research would thus include an assessment of these models on the same footing together with a thorough investigation of the impact of the power transformation of the realized volatility series.

Chapter 3

Learning from Micro-level Expert Forecasts: Real-time Data, Regression Trees, and Bagging

Simon D. Knaus
Fabian Krüger

Abstract

While macroeconomic survey forecasts are widely available at the level of individual experts, it is not clear how to optimally combine a set of forecasts to a "consensus" prediction. This is mainly due to the characteristics of the data, such as the large-dimensional predictor space, many missing values, and potential individual and aggregate level biases of the survey forecasts. We argue that regression trees are very well adapted to these features and propose to use them as a novel forecast combination device. Our empirical analysis of data from the Philadelphia Fed's Survey of Professional Forecasters demonstrates that in combination with bagging, tree-based forecast combination outperforms equally weighted combination for the majority of time series and forecast horizons.

JEL: C14, C23, C43, C53

Keywords: Macroeconomic Survey Data, Survey of Professional Forecasters, Bagging, Regression Trees, Real-time Data

3.1 Introduction

The combination of point forecasts has been an econometric success story initiated already by Bates & Granger (1969). *Ex ante*, a researcher who does not know which individual forecast model will perform well in the future is most often well advised to use combination strategies which provide a hedge against idiosyncratic model failure. See Clemen (1989) and Timmermann (2006) for reviews of the literature.

While the literature on forecast combination is extensive, the vast majority of contributions focuses on the case where a set of econometric model forecasts is considered for combination. This case is "well-behaved" in the sense that the individual models are controlled by the econometrician. Hence, it can be ensured that the models' forecast-generating processes are understood and forecasts from each model are available in each period. In sharp contrast to this setting, users of individual-level expert forecasts collected in surveys typically do not know how these forecasts were obtained, and must deal with a large fraction of missing values. Nevertheless, surveys such as the Survey of Professional Forecasters, the Livingston Survey, and several commercial sources, have recently received much attention in the forecasting literature. This attention has been reinforced by promising results on the accuracy of survey forecasts of inflation, as compared with a large set of time series models (Ang, Bekaert & Wei 2007, Faust & Wright 2011).[1]

Users of surveys face the question of how to aggregate the set of available expert forecasts into a single number, often called the "consensus forecast". Forecast combination in surveys poses some specific challenges, due to the frequent entry and exit of participants and the large cross-sectional dimension of the relevant data sets. A simple response to these challenges is to use a summary statistic from the cross-section of forecasters at a given point in time, such as the mean or median, as a consensus forecast. These summary statistics are parameter-free and have a proven track record in general; see e.g. Smith & Wallis (2009). On the negative side, they neglect information at the level of individual forecasters: they are invariant under permutations of "who says what". Most existing alternatives which

[1] Independently of the question whether survey forecasts are accurate or not, survey data can naturally be used to test hypotheses about expectations formation. See Pesaran & Weale (2006) for a review of this literature.

tackle this shortcoming (Capistrán & Timmermann 2009, Poncela, Rodríguez, Sánchez-Mangas & Senra 2011) use imputation techniques to solve the missing data problem, and then work with the preprocessed data. This approach is problematic for two reasons: First, little theoretical or empirical guidance is available for performing imputation in this context. Second, imputation in practice inevitably leads to a reduction in the cross-sectional variance of individual-level forecasts, which is at odds with the major motivation for looking at individual-level information in the first place: its heterogeneity.

In contrast, this paper seeks to overcome these problems by exploring the use of regression trees (Breiman, Friedman, Olshen & Stone 1984) for forecast combination in surveys. Trees are very popular in the statistical learning literature but largely unexplored by the econometric community. They are very well adapted to the characteristics of survey data, for three reasons: First, by using so-called "surrogate splits" they can deal with the missing data problem without performing explicit imputation. If a particular expert forecast is unavailable, the pool of available forecasters is "screened to find a surrogate" that mimics the historical predictions of the missing expert as closely as possible. Second, by means of a built-in variable selection they can deal with the large dimension of the predictor space spanned by the individual survey participants. Third, trees can account for the (potentially) complicated biases of survey forecasts at the individual and aggregate level. Trees are particularly promising in conjunction with some robustification device which remedies the sensitivity of the base variant to small changes in the data. We use bagging (Breiman 1996) for this purpose. Bagging is based on the idea of re-estimating a model for several bootstrap resamples of the training data and then considering the average of the resulting set of predictions.

Few of the combination approaches considered in the literature are able to handle missing values in a natural way. One of these approaches is analyzed by Capistrán & Timmermann (2009). The authors consider a selection scheme which picks the best forecaster in the training sample.[2] Including this scheme in our analysis helps to disentangle the potential benefits of individual-level data *per se* and the benefits of trees as a method. Similarly to trees, however, the base variant of this scheme is unstable since the selection of a particular

[2]Since each forecaster's performance is evaluated only on the basis of the available training-sample observations of this forecaster, this scheme can also be used in the presence of missing values.

forecaster may be due to a small number of training observations. We show that bagging leads to large improvements in the performance of this scheme, rendering it a viable alternative to existing combination methods considered in the literature.

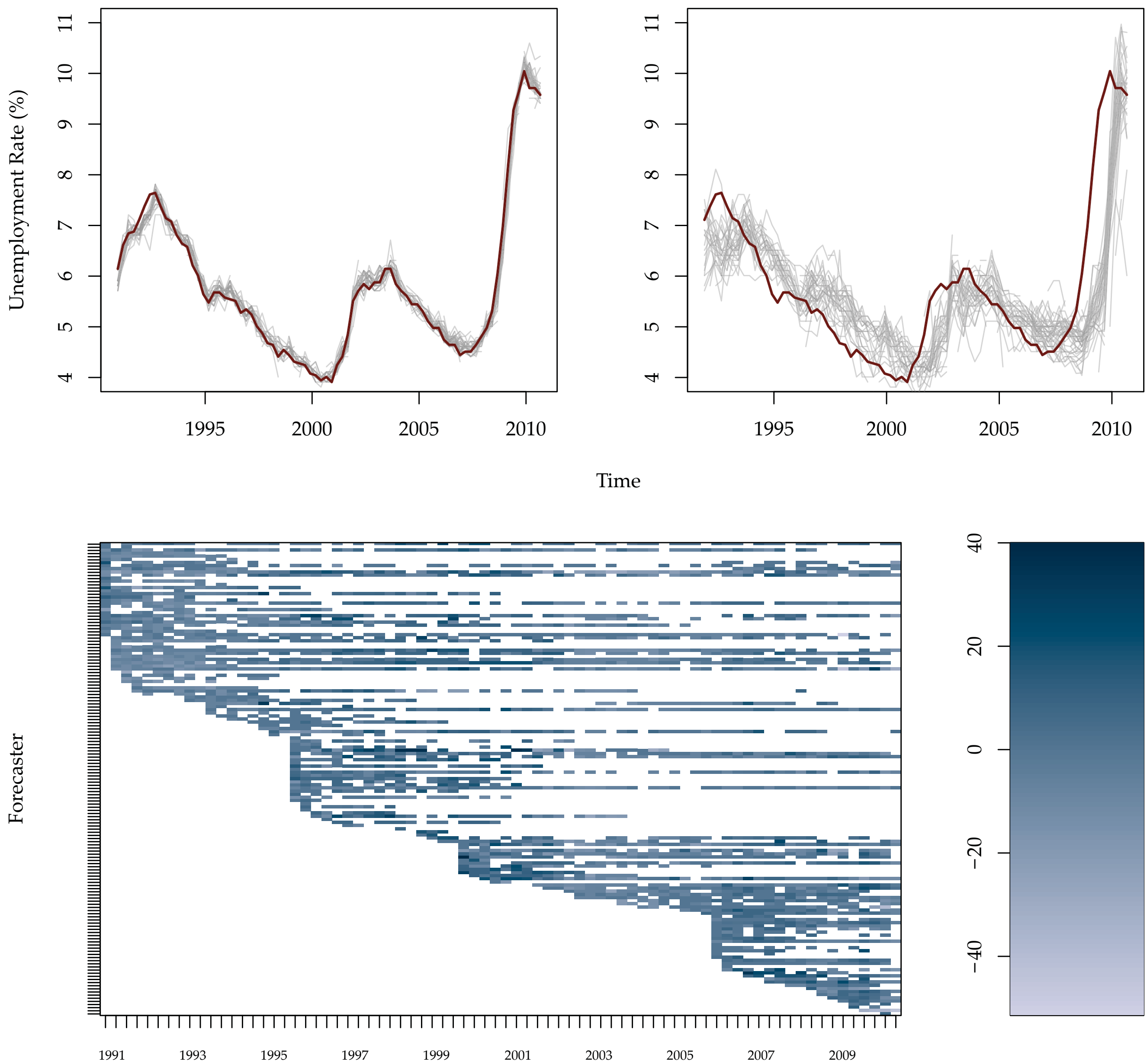
It is important to note that designing sophisticated combinations of individual-level survey forecasts is futile unless there is *informative* heterogeneity in the forecasts. The literature on expectations formation suggests a number of sources for "structural", and thus informative, heterogeneity. Assuming that experts solve a well-defined decision problem (i.e., minimizing expected loss) when deciding on their forecast, disagreement may arise if forecasters have different loss functions, different information sets, or different beliefs given a particular information set. See Pesaran & Weale (2006, Section 2) for an instructive discussion and Patton & Timmermann (2010) for a structural model of forecaster disagreement.

To illustrate forecast heterogeneity in the data, the two upper panels of Fig. 3.1 show forecasts of the US unemployment rate, a nowcast (left panel) and four quarters (right panel) into the future. The graphs illustrate that especially at longer horizons, there is substantial disagreement among forecasters. The lower panel of Fig. 3.1 takes another look at unemployment forecasts.[3] The camouflage-type look of the graph illustrates substantial forecast heterogeneity at any given point in time. Furthermore, the large number of white spots on the panel illustrates the quantitative importance of the missing-data problem.

The remainder of this paper is organized as follows. Section 2 describes trees and bagging. Section 3 describes the strict real-time procedure we use in order to realistically estimate and evaluate all models. Section 4 presents our empirical results, and Section 5 concludes. Additional results and technical details are available in the Appendix.

[3] In contrast to the upper panel, the lower panel of Fig. 3.1 refers to forecasted *changes* of the unemployment rate.

Figure 3.1: Unemployment forecasts



Unemployment Forecasts. The top left (top right) panel shows the zero quarter (four quarters) ahead forecast of the unemployment rate as one gray line per forecaster with the true ex post value in bold red. The lower panel shows one quarter ahead forecasts of quarterly *changes* in the unemployment rate. Each row is a forecaster (i), each column is a date (t). The color, as coded by the scale at the right hand side, at the intersection (i, t) corresponds to the forecast of i submitted at t. Non-colored dots (white) are missing values.

3.2 Methods

To accommodate the particularities of survey data that warrant particular attention we employ two combinations schemes: regression trees and a simple selection scheme. Both methods are fragile in their base variant but will be robustified by bootstrap aggregating. We introduce all these methods in the following.

3.2.1 Regression Trees

Regression trees have a long history: They were first mentioned in Morgan & Sonquist (1963) to uncover masking and cope with interaction in classical surveys, e.g., wage surveys. The influential work of Breiman et al. (1984) contained the first unified presentation of classification and regression trees. Trees have enjoyed great popularity in computational statistics (Hastie et al. 2009) and in the machine learning community (Ripley 2008) and have been extended along various directions. Some statistical properties of regression trees have been established in Breiman et al.'s (1984) original work and later in Gey & Nedelec (2005). Trees have been successfully applied to various problems, e.g., GARCH modeling (Audrino & Bühlmann 2001), the shape of the risk-return premium (Rossi & Timmermann 2010), structural break detection (Rea, Reale, Cappelli & Brown 2010), and volatility surfaces (Audrino & Colangelo 2010).

The basic mechanism of a regression tree follows from its name: it is a regression (an estimate of a conditional expectation) by means of binary recursive partition (which can be illustrated by an abstract picture of a tree). A tree is best thought of as a set of locally constant predictions. Let X_t be a p-dimensional vector of predictors in $\mathbb{R}^p$, Y_t the dependent variable in $\mathbb{R}$, and $P_1, \ldots, P_m$ a partition[4] of the predictor space $\mathbb{R}^p$. The tree's prediction $\widehat{tr}(X_t)$ is given as the average over all values of Y in one particular set P_k, i.e.,

$$\widehat{tr}(X_t) = \frac{1}{|\{X_j : X_j \in P_k\}|} \sum_{j:X_j \in P_k} Y_j \qquad \text{where } X_t \in P_k.$$

The apparent pivotal part is the partition P since the prediction of a tree crucially depends on how $\mathbb{R}^p$ is split into cells P_k. The partition P is determined in a binary recursive way: binary since each split determines two cells and recursive as the next splits are determined within the two previous cells. The splits can thus occur at any value of any variable $(X^1, \ldots, X^p)' \in \mathbb{R}^p$ and are determined such that the training sample mean squared error (MSE) of the tree's prediction $\widehat{tr}(X_t)$ decreases maximally. This recursive partitioning[5] continues until the tree is maximal (only one observation in one terminal leave) and is then pruned by deleting leaves of the maximal tree by cost-complexity pruning.[6] For an intuitive illustration of the mechanics and visual interpretation of a regression tree we refer to Fig. 3.2. This illustration is of course of exemplary nature: It nicely illustrates the tree's ability to pick up non-linearities. A complete formal treatment of how a tree is built in our particular application is relegated to Appendix C.2.1.

An important feature of this greedy recursive partition mechanism is that trees act as a variable selection and regression device at the same time. The recursive binary partition scheme is crucial: It remedies the curse of dimensionality and yet still allows for regressing non-linear structures. Exploiting this feature one can – unlike in traditional multi-dimensional (non-parametric) regression – still tackle problems with a high-dimensional predictor space.

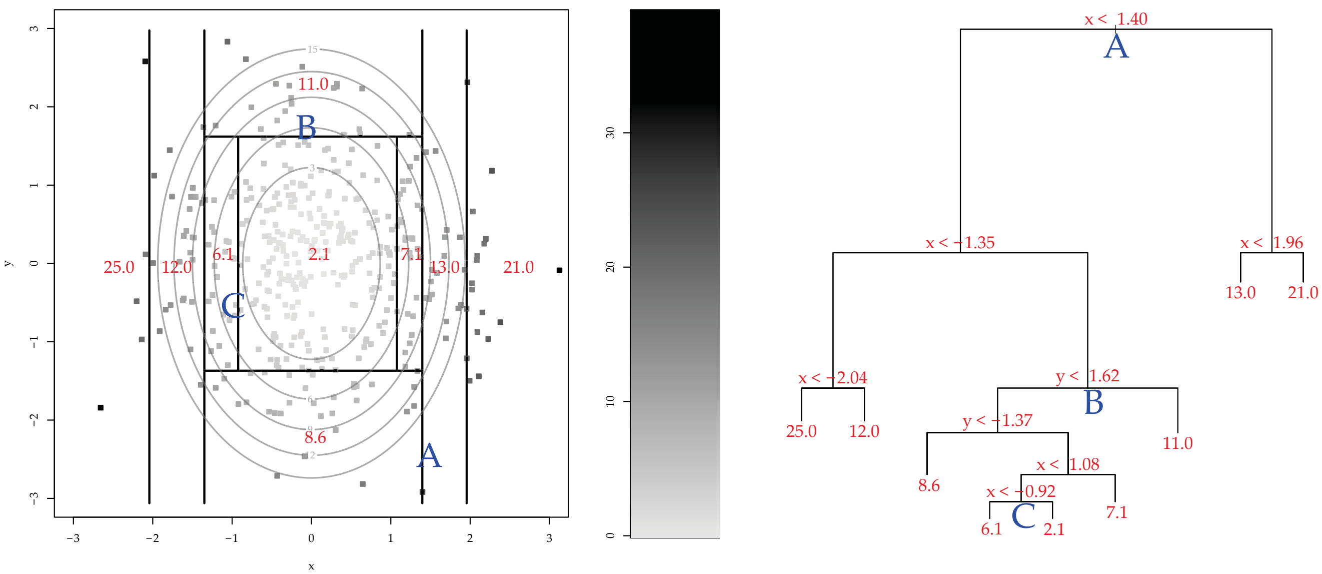
A further unique feature of trees is surrogate splits. Surrogate splits allow observations to descend the tree further down (to ultimately obtain a prediction) even if there are missing values by determining a replacement split which optimally mimics the splitting as if the split variable was present. For each split a tree with the dependent variable 'go left'

[4]In a strict mathematical sense, a partition $P = \{P_1, \ldots, P_m\}$ of a set S is complete ($S = \cup_{i \in \{1,\ldots,m\}} P_i$) and non-overlapping ($P_i \cap P_j = \emptyset$ for any $i \neq j$).

[5]Observe that the partition P obtained in this recursive manner consists of halfspaces and thus leads to convex sets with boundaries being hyperplanes perpendicular to the axes.

[6]Cost-complexity pruning is the method which is employed to find the optimal leaves to delete; a discussion is found in the Appendix.

Figure 3.2: Regression tree



In this toy example we depart from $Z_i = 4\,X_i^2 + 2\,Y_i^2 + \varepsilon_i$ and estimate $\mathrm{E}(Z|X,Y) = f(X,Y) = 4\,X^2 + 2\,Y^2$ by means of a regression tree. The true function $f(X,Y)$ is a parabola with the minimum in zero, growing faster in the direction of X than Y. The left panel shows the values of Z_i as color-coded by the bar in the mid panel and the corresponding values of X_i (abscissa) and Y_i (ordinate). The contour plot (the concentric ellipses) indicates the isoquants of $f(X,Y)$. The bold lines correspond to the binary recursive partition as obtained by a regression tree illustrated in the right panel (containing the mean for each cell of the partition). The right hand side panel displays the recursive partition illustrated in a tree manner. The split values as well as the averages in each leave are shown. The splits (rhs) can be identified with the bold lines (lhs), and so can the cell means. Splits A to C are labeled explicitly on both sides.

and 'go right' is fitted to the data at this particular split in question. This in-situ tree is then used to predict the split if the split variable is missing. For a complete treatment of surrogate splits the reader is referred to Breiman et al. (1984).

3.2.2 Previous Best Selection

The previous best scheme simply uses the prediction of the forecaster who attained the lowest MSE in the training sample. This idea has already been introduced by Capistrán & Timmermann (2009) and is remotely related to the idea of combining forecasts based on their precision (Timmermann 2006, Section 2.3), i.e., the inverse of forecasts' covariance-matrix, but implemented more directly since the estimation (missing values) and inversion (almost singular) of the covariance-matrix are typically non-trivial. A valuable feature of the selection approach is that it is not affected by missing values since the MSE of each forecaster is simply calculated over the training-sample observations for which predictions of this forecaster are available. For a formally rigorous account of how the selection approach works in our application we refer to Appendix C.2.2. A related idea is presented in Gupta & Wilton (1988): the selection is implicit in the determination of the combination weights. However, the process of determining the weights relies on the computation of the eigen-values of the so-called odds-matrix, which is a square matrix where the dimension equals the number of forecasters. The large fraction of missing values (although this method could be modified to deal with missing values) together with the large number of forecasters renders this method inappropriate for our purposes.

3.2.3 Bagging

Bootstrap aggregating, in short bagging, was introduced in Breiman (1996) and has been successfully applied to a variety of problems. In essence, bagging consists of resampling and averaging models. Adopting the language of the statistical learning community we depart from a training sample $\mathcal{L}$ consisting of observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where typically X_t is a multi-dimensional variable and Y_t a one-dimensional, continuous outcome. Let $\hat{f}_{\mathcal{L}}$ be a statistical learner[7] trained on a sample $\mathcal{L}$. Bagging then consists of drawing B bootstrap replicates $\{\mathcal{L}_1, \ldots, \mathcal{L}_B\}$ each of length n and estimating the learner on each $\mathcal{L}_b$, resulting in a set of trained learners $\{\hat{f}_{\mathcal{L}_1}, \ldots, \hat{f}_{\mathcal{L}_B}\}$. The bagged prediction $\hat{f}^B$ is then the average over the learners trained on the bootstrap replicates of $\mathcal{L}$, i.e.,

$$\hat{f}^B(X) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_{\mathcal{L}_b}(X).$$

Bagging proves particularly useful in case of unstable predictors; unstable in the sense that a small perturbation in one of the predictors can lead to a significant change in the

[7] A statistical learner is a very general term. Consider the simple case of OLS regression $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. The learning procedure then consists of estimating (β_0, β_1) with $(\hat{\beta}_0, \hat{\beta}_1)$ on $\mathcal{L}$ and consequently $\hat{f}_{\mathcal{L}}$ is given by $\hat{f}_{\mathcal{L}}(X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Equivalently, the learning process in the case of a regression trees consists of determining the partition $P_1, \ldots, P_m$ on the learning sample $\mathcal{L}$.

predicted outcome. Consequently, $\hat{f}_{\mathcal{L}_b}$ is expected to differ substantially across the different replicates $\mathcal{L}_b$. A more detailed account of the merits of bagging is given in Breiman (1996) and extended in Bühlmann & Yu (2002).

Recently, bagging has been applied to the prediction of consumer price inflation (Inoue & Kilian 2008), unemployment (Rapach & Strauss 2010), volatility (Hillebrand & Medeiros 2010) and interest rates (Audrino & Medeiros 2011).

We apply bagging to both the regression trees and the previous best forecast scheme; we generate the bootstrap replicates with the stationary bootstrap (Politis & Romano 1994) and an expected block length of two years. In the former scheme the bagged predictor is an average of the B predictions of B (different) trees, in the latter scheme it is the average of the B best forecasters determined in each of the samples $\mathcal{L}_b$. We choose 100 bagging iterations for all base variants.[8]

3.2.4 Discussion of Methods

It is important at this point to review and contrast the proposed methods. Regression trees, in their base variant as well as their bagged variant, will use the forecasts as predictors and the corresponding real-time realization of a macroeconomic indicator as dependent variable. By regressing the realization on the forecast we can potentially eliminate the bias inherent to the submitted forecasts. In contrast, the previous best selection scheme only relies on past real-time realizations through the selection of the best performing forecaster, but will eventually still give one specific forecaster's submission verbatim and thus lacks the option of accounting for a potential bias at the individual level.

Despite the absence of a bias correction in the previous best scheme it is instructive to compare results of this method to the outcome of regression trees: It should help to

[8] The number of bagging iterations, if large enough, should not be considered a tuning parameter, see Bühlmann (2004)

(empirically) disentangle a potential benefit due to considering micro-level information (which both methods feature) and due to bias correction (which only trees feature).

3.3 Data and Procedures

3.3.1 Data

Our proposed methods apply to any kind of forecast data which exhibit the features outlined above. For illustration purposes, we choose the Survey of Professional Forecasters (SPF) for our analysis, which has received considerable attention in recent years.

The SPF was initiated and administered by the National Bureau of Economic Research (NBER) and the American Statistical Association (ASA) in 1968 and taken over by the Federal Reserve Bank of Philadelphia (Philadelphia Fed) in 1990. The survey form is sent out quarterly to professionals in different fields (academia, industry, banking, etc.) and covers over 30 macroeconomic variables, some of which are included since the inception of the survey. The forecasters are invited to submit their forecast for the current quarter (nowcast) as well as the next four quarters (one to four quarters ahead forecasts) in addition to a forecast of the current and next year's annual development for each of the variables contained in the survey. Hereby all forecasters dispose of the latest preliminary release of the last quarter's realization. The submitted point forecasts are available at micro-level identified by an anonymized ID assigned to each forecaster. Unfortunately the consistency of the forecasters' IDs cannot be guaranteed for the period when the SPF was administered by NBER and ASA requiring us to only use survey data sent out after 1990Q3. From 1990Q3 onwards there are 160 IDs in the dataset whose point forecasts of the quarterly data we use in our analysis. A complete discussion of the timing, variables, transformations, and horizons is found in SPF (2010).

For our analysis we consider the series[9] displayed in Tab. 3.1. As can be inferred from Tab. 3.1 the fraction of missing invidual-level forecasts is high for each of the series rendering methods that depend on complete datasets unfeasible. Since all methods outlined in Section 3.2 are able to deal with missing values, the high fraction of non-submissions (or equivalently, the frequent entry and exit of experts) does not pose a problem per se.

Aspiring to be maximally realistic in our application of forecast combination we pretend "to walk in the forecasters' shoes" and use real-time vintages, i.e. all forecasts could have been calculated in the very same way at any given point in time. Real-time data is provided by the Philadelphia Fed, too, and is updated quarterly to reflect ex post revisions of a macroeconomic datum. The aggregation from monthly to quarterly data of the real-time data is carried out in accordance with SPF standards (Stark 2010), i.e., quarterly data is calculated from three month averages of the original series. If necessary, the resulting quarterly series are transformed to stationarity thereafter. See Tab. 3.1 for the transformations we use. In the next subsection, we provide a more formal description of our real-time procedure.

[9] The fact that some of the variables surveyed by the SPF suffer from changes in the base year and the requirement that most of the time series need to be transformed, leaves us with the series of Tab. 3.1. These coincide with the series analyzed in Poncela et al. (2011) with the exception of US treasury bond yields (TBOND).

Table 3.1: Data description

			Descriptive Statistics					
Series	Description	Missing (%)	ACF(1)	ACF(5)	Mean	Std	LB(5)	ADF
BOND	Avg. bond yield for Moody's AAA corporate bond (none, B)	76.1	0.15	−0.17	−0.06	0.29	10.13*	−5.13**
CPI	Annual Inflation rate (saar, A)	74.7	0.15	0.04	2.58	2.02	3.95	−4.02**
HOUSING	Level of Housing Starts (saar, C)	74.1	0.17	0.11	0.06	26.23	8.13	−3.84**
NGDP	Level of Nominal GDP (saar, C)	74.2	0.50	0.08	4.75	2.75	43.15***	−3.46*
TBILL	Three month treasury bill rate (none, B)	74.8	0.62	−0.03	−0.09	0.45	64.70***	−3.93**
TBOND	Ten year treasury bond rate (none, B)	76.1	0.17	−0.28	−0.07	0.38	16.43***	−5.42**
UNEMP	Unemployment rate (sa,B)	73.9	0.73	0.12	0.05	0.30	86.21***	−3.36*

Column two contains a description of the variable with the transformation in parentheses. The first transformation (sa: seasonally adjusted; ar: annual rate) is predetermined by the Philadelphia Fed, the second transformation (A: no transformation; B: first differences; C: annualized growth rate, $\left((x_t/x_{t-1})^4 - 1\right) \cdot 100$) is to have stationary data. Column three contains the percentage of missing individual-level forecasts (non-submissions) for each of the series. Descriptive statistics (Mean, Standard Deviation, autocorrelation at lags 1 and 5) for each of the transformed series is reported along with Ljung-Box statistics of order 5 and an Augmented Dickey-Fuller statistic (***: p-value < 0.01, **: p-value < 0.05, *: p-value < 0.1) in columns 4 to 9.

3.3.2 Procedures

Our analysis is based on a strict real-time approach. Denote by Y the stationary quarterly variable of interest. For each quarter t, we have a set of SPF expert forecasts $\{\hat{Y}^i_{t,h}\}_{i \in \mathcal{E}_{t,h}}$, where $\hat{Y}^i_{t,h}$ is the h-step ahead forecast of expert i and $\mathcal{E}_{t,h}$ is the set of individuals with non-missing entries. Apart from these survey data, we select the historical information set of the target variable Y that would have been realistically available to the SPF participants. Since the SPF forecasts are collected in the middle of each quarter, the current quarter's value Y_t is unavailable at t.[10] For this reason, a nowcast (a forecast with horizon $h = 0$)

[10]See Stark (2010) and SPF (2010) for additional information on the timing of the SPF.

is generally a non-trivial exercise. Furthermore, past data of Y may be revised as time progresses; see Croushore (2006). To account for this setting, denote by Y^t_{t-j} the value of Y_{t-j}, using the most recent vintage available at t.[11]

In order to assess the accuracy of a forecasting method, we compute its mean squared prediction error (MSPE) as

$$\text{MSPE} = \frac{1}{L} \sum_{t=T}^{T+L-1} (Y_t^{\infty} - \tilde{Y}_{t-h,h})^2,$$

where L is the size of the evaluation sample, the notation Y_t^{∞} suggests that the latest available vintage of all realized data is used for evaluation[12] and $\tilde{Y}_{t-h,h}$ denotes a generic h-step ahead forecast made at date $t-h$, henceforth referred to as the "origin date". Note that MSPE is simply the out-of-sample counterpart of MSE, which we use as an in-sample criterion to estimate all forecasting methods.

The micro-level methods introduced in the last section aim at exploiting information at the level of individual survey participants. We refer to the tree- and selection methods and their bagged variants as *Tree*, *TreeBagg*, *Slct* and *SlctBagg* in the following. In order to put their performance in perspective, we compare them to three simple but effective alternative methods. First, equally weighted (*EW*) combination of expert forecasts serves as a natural benchmark method with a very good track record for forecast combination in general (Timmermann 2006, Smith & Wallis 2009). Second, the bias correction proposed by Capistrán & Timmermann (2009), henceforth *CapTim*, is a linear regression of the realized value of Y on its mean survey forecast and an intercept. Third, we use an autoregressive (*AR*) model whose lag length is adaptively chosen via the Akaike (1970) information criterion.

The forecasting methods used in this paper are based on different information sets comprising expert forecasts and/or historical realizations of Y. While *AR* is based on real-time

[11] As described above, Y^t_{t-j} is only available for $j = 1, 2, \ldots$.

[12] In practice we use the data vintage of 2011Q1.

data of Y only, *EW* uses only the current average survey prediction. *CapTim* uses historical realizations of Y and historical average survey predictions. The micro-level methods analyzed in Section 3 use historical realizations of Y and the full set of historical individual-level forecasts. We formally describe these information sets in Tab. 3.2 and Tab. 3.3 below. In the latter table we differentiate between the information set used for estimation (i.e., to determine a mapping from signals to predictions) and the information set used for out-of-sample prediction.

In practice, at each origin date t, we estimate each forecast combination scheme using a rolling window of 40 observations. We then compute predictions based on each scheme. This is repeated for different forecast horizons $h = 0, 1, 2, 3, 4$. Since all combination schemes we consider are in the spirit of "direct regression models" (Marcellino, Stock & Watson 2003, 2006), we re-estimate each scheme for each forecast horizon h. Once this is complete, we move to the next forecast origin date $t + 1$.

Table 3.2: Notation

Symbol	Interpretation
$\hat{Y}^i_{t,h}$	Forecast of expert i, with origin date t and target date $t + h$
$\mathcal{E}_{t,h}$	Set of experts who submit an h-step ahead forecasts at t
$\hat{Y}_{t,h} = \frac{1}{\lvert\mathcal{E}_{t,h}\rvert} \sum_{i \in \mathcal{E}_{t,h}} \hat{Y}^i_{t,h}$	Equally-weighted (*EW*) combination of h-step ahead forecasts
$\mathcal{F}^{\text{ew}}_t = \{\hat{Y}_{t-j,h}\}_{j=0,1,\ldots}$	Set of historical *EW* forecasts
$\overline{\mathcal{E}}_{t,h} = \cup_{j=0,1,\ldots} \mathcal{E}_{t-j,h}$	Set of experts who *ever* submitted an h-step ahead forecast *until t*
$\mathcal{F}^{\text{mic}}_t \equiv \{\hat{Y}^i_{t-j,h}\}_{i \in \overline{\mathcal{E}}_{t,h}, j=0,1,\ldots}$	Set of historical micro-level forecasts
Y^t_{t-j}	Time-t vintage of variable Y_{t-j}
$\mathcal{F}^{\text{dat}}_t \equiv \{Y^t_{t-j}\}_{j=1,2,\ldots}$	Most recent vintages of historical data, as of t

Notation summarizing real-time information used by the different forecast combination schemes.

Table 3.3: Relevant information sets

	Information Set	
Method	Estimation	Prediction
AR(p)	$\mathcal{F}_t^{\text{dat}}$	$\{Y_{t-p+1}^t, \ldots, Y_{t-1}^t\}$
EW	–	$\hat{Y}_{t,h}$
CapTim	$\mathcal{F}_t^{\text{dat}}, \mathcal{F}_t^{\text{ew}}$	$\hat{Y}_{t,h}$
Slct, *SlctBagg*,	$\mathcal{F}_t^{\text{dat}}, \mathcal{F}_t^{\text{mic}}$	$\{\hat{Y}_{t,h}^i\}_{i \in \mathcal{E}_{t,h}}$
Tree, *TreeBagg*	$\mathcal{F}_t^{\text{dat}}, \mathcal{F}_t^{\text{mic}}$	$\{\hat{Y}_{t,h}^i\}_{i \in \mathcal{E}_{t,h}}$

Information sets used by different forecast combination schemes. For simplicity, the information sets shown here are comprehensive, i.e. the forecasting methods do not necessarily use all of the information in these sets.

3.4 Results

3.4.1 Main Results

The main results of our forecasting study are displayed in Tab. 3.4. Our findings can be summarized as follows. First, we provide evidence that equally weighted combination of expert forecasts can be improved upon. Bagged regression trees achieve lower MSPEs than *EW* in 23 out of the 35 forecast comparisons (= 7 variables × 5 horizons) shown below. Bagged forecaster selection, the linear bias correction scheme proposed by Capistrán & Timmermann (2009, *CapTim*), and trees also perform relatively well, achieving lower MSPEs than *EW* in 14, 15 and 18 comparisons, respectively. To shed further light on the relative performance of the alternative methods, we use the Superior Predictive Ability (SPA) test by Hansen (2005). The null hypothesis of this test is that a given benchmark method is *not* dominated by any competitor in terms of MSPE. Hence, a benchmark method is successful in the sense of the SPA test if its "undominatedness" can rarely be rejected across the 35 forecast comparisons. Using a 5 % significance level and varying benchmark methods, Tab. 3.4 shows that *CapTim* incurs the smallest number of rejections when used as a benchmark, followed by *TreeBagg*, *EW*, *Tree*, *SlctBagg* and *Slct*.

Second, the performance of both micro-level combination methods can be ameliorated via bagging, more pronounced however for *Slct* where the improvement is often by large MSPE margins. Using one-sided Diebold & Mariano (1995) tests, the bagged versions of *Tree* and *Slct* significantly outperform their base variants in 11 and 15 comparisons at the 10% level (see Table C.1 in the Appendix). For both combination schemes, bagging effectively cures the ill-conditionedness of the base variant, which implies that the prediction of the base learner is not robust to small perturbations of the training data. In the case of *Tree*, ill-conditionedness arises from the binary (hard) splits partitioning the (large-dimensional) predictor space. In the case of *Slct*, instabilities result from the fact that the identity of the forecaster with the lowest MSPE is likely to differ across different sub-samples of the training sample. The less pronounced improvement of *Tree* can be explained by the fact that already the base variant is robustified in the sense that redundant splits are likely to be deleted during pruning (see Sections 3.2.1 and C.2.1).

Although the focus of our paper is on the optimal use of survey forecasts, rather than the relative merits of survey forecasts and other methods, looking at the performance of a simple time series model can be instructive. We therefore also compare *EW* combination of survey forecasts to a univariate autoregressive (*AR*) model whose lag length is adaptively selected via the Akaike information criterion for each training period. We find that surveys perform extremely well at horizon 0; the relative MSPEs of *AR* relative to *EW* range from 1.14 (HOUSING) to 4.89 (TBILL). These findings can partly be explained by intra-quarter information on the target series which is available to survey participants but not to the *AR* model.[13] Note, however, that surveys also clearly outperform *AR* for NGDP where this argument does not apply. These findings regarding the very good "nowcasting" performance of the mean SPF forecast are in line with Stark (2010) who considers different variables and/or evaluation periods than we do. Regarding the relative performance of

[13]When SPF participants submit their forecasts in the middle of a quarter, roughly half of the daily rates entering the quarterly variable to be forecast are already available in the case of the interest rate series. Values for the quarter's first month are released around the survey date for CPI and HOUSING, while first-month values of UNEMP are already available to survey participants. In contrast, no preliminary estimate of the current quarter's value is available in the case of NGDP. See Stark (2010) for further information and references.

surveys and *AR* at longer horizons, our results are more mixed: While surveys tend to do better than *AR* for CPI, NGDP, TBILL and UNEMP, the reverse is true for the other series.

Table 3.4: Full evaluation sample

		Forecast Combination Methods (in SPA set)						
		Tree	*TreeBagg*	*Slct*	*SlctBagg*	*CapTim*	*EW*	*AR*
BOND	0	0.96 (0.03)	0.67 (0.84)	1.51 (0.15)	0.98 (0.49)	0.92 (0.05)	1.00 (0.00)	1.81
	1	0.87 (0.94)	0.90 (0.54)	1.34 (0.00)	1.12 (0.04)	0.95 (0.17)	— (0.19)	0.82
	2	0.82 (0.57)	0.75 (0.83)	0.96 (0.26)	0.98 (0.26)	0.77 (0.80)	— (0.12)	0.71
	3	0.78 (0.65)	0.98 (0.10)	1.29 (0.03)	1.10 (0.04)	0.98 (0.15)	— (0.05)	0.79
	4	0.78 (0.93)	0.97 (0.04)	1.07 (0.03)	1.01 (0.03)	0.84 (0.14)	— (0.04)	0.78
CPI	0	2.11 (0.01)	1.36 (0.08)	1.01 (0.56)	0.94 (0.90)	0.91 (0.79)	1.00 (0.45)	2.73
	1	1.08 (0.17)	1.04 (0.56)	1.03 (0.40)	1.01 (0.67)	1.02 (0.76)	— (0.97)	1.10
	2	1.02 (0.11)	0.93 (0.92)	1.06 (0.22)	1.04 (0.30)	1.04 (0.17)	— (0.36)	1.04
	3	1.54 (0.23)	0.98 (0.94)	1.06 (0.26)	1.07 (0.21)	1.06 (0.51)	— (0.71)	1.03
	4	1.03 (0.73)	1.03 (0.55)	1.04 (0.48)	1.05 (0.47)	1.07 (0.31)	— (0.92)	1.04
HOUSING	0	0.96 (0.12)	0.70 (0.80)	1.33 (0.01)	0.93 (0.01)	1.13 (0.04)	1.00 (0.09)	1.14
	1	0.83 (0.10)	0.66 (0.93)	1.08 (0.19)	0.99 (0.27)	1.03 (0.08)	— (0.20)	0.79
	2	0.74 (0.83)	0.74 (0.90)	0.73 (0.80)	0.78 (0.41)	1.65 (0.12)	— (0.09)	0.65
	3	0.71 (0.05)	0.64 (0.96)	3.04 (0.10)	1.37 (0.10)	1.41 (0.21)	— (0.05)	0.61
	4	0.78 (0.96)	0.81 (0.58)	2.13 (0.19)	1.20 (0.31)	1.61 (0.08)	— (0.24)	0.68
NGDP	0	2.13 (0.01)	1.87 (0.06)	1.41 (0.02)	1.21 (0.03)	1.29 (0.08)	1.00 (0.70)	2.67
	1	1.65 (0.07)	1.54 (0.03)	1.13 (0.08)	0.94 (0.83)	1.34 (0.02)	— (0.35)	1.61
	2	1.44 (0.07)	1.34 (0.04)	1.02 (0.52)	0.99 (0.98)	1.34 (0.04)	— (0.44)	1.29
	3	1.24 (0.01)	1.28 (0.00)	0.95 (0.82)	0.97 (0.48)	1.25 (0.01)	— (0.24)	1.10
	4	1.08 (0.04)	1.07 (0.07)	1.03 (0.05)	0.99 (0.67)	1.12 (0.01)	— (0.45)	1.08
TBILL	0	0.57 (0.57)	0.99 (0.04)	0.90 (0.00)	0.75 (0.00)	0.53 (0.63)	1.00 (0.04)	4.89
	1	1.18 (0.00)	0.66 (0.60)	1.53 (0.01)	1.08 (0.00)	0.67 (0.40)	— (0.02)	1.10
	2	0.96 (0.42)	0.92 (0.60)	1.19 (0.15)	1.15 (0.05)	0.85 (0.99)	— (0.37)	0.96
	3	0.94 (0.45)	0.83 (0.93)	1.48 (0.03)	1.26 (0.05)	0.88 (0.85)	— (0.35)	1.09
	4	1.25 (0.06)	0.87 (0.87)	1.33 (0.01)	1.07 (0.18)	1.03 (0.47)	— (0.64)	1.02
TBOND	0	0.95 (0.23)	0.81 (0.90)	1.93 (0.04)	1.50 (0.03)	0.92 (0.39)	1.00 (0.07)	2.85
	1	0.83 (0.81)	0.84 (0.88)	0.97 (0.44)	0.97 (0.51)	1.09 (0.22)	— (0.35)	0.85
	2	1.12 (0.03)	0.88 (0.14)	1.34 (0.01)	1.17 (0.02)	0.79 (0.92)	— (0.01)	0.84
	3	0.77 (0.99)	1.00 (0.25)	0.95 (0.13)	1.01 (0.02)	1.09 (0.14)	— (0.11)	0.83
	4	0.87 (0.99)	0.98 (0.17)	1.17 (0.17)	1.11 (0.22)	1.02 (0.26)	— (0.35)	0.85
UNEMP	0	2.03 (0.00)	1.90 (0.05)	1.20 (0.02)	1.03 (0.09)	0.88 (0.98)	1.00 (0.17)	3.71
	1	1.82 (0.03)	1.30 (0.05)	1.21 (0.03)	1.07 (0.06)	0.94 (0.79)	— (0.35)	1.82
	2	1.44 (0.10)	1.05 (0.10)	0.83 (0.86)	0.89 (0.55)	1.04 (0.05)	— (0.40)	1.42
	3	1.23 (0.12)	0.99 (0.76)	1.12 (0.11)	1.02 (0.66)	1.04 (0.08)	— (0.84)	1.14
	4	0.94 (0.44)	1.03 (0.10)	0.86 (0.83)	0.90 (0.58)	0.97 (0.35)	— (0.26)	0.99

Forecast evaluation from $T_0 + h$ to T_1 where T_0 = 2000Q4, T_1 = 2010Q4, and $h = 0, \dots, 4$ ($41 - h$ quarterly evaluation points). All combination methods use a rolling training sample of 40 observations.[14] Numbers in parantheses are p-values of the Superior Predictive Ability (SPA) test by Hansen (2005) with corresponding method used as benchmark; we use a mean block length of 8 observations and 10.000 replications in the bootstrap implementation. All other numbers are MSPEs, relative to the mean SPF prediction (*EW*).

3.4.2 Discussion

In order to better understand the relative strengths and weaknesses of the different forecast combination methods and eventually explain their relative performance we resort to a well known decomposition. If a probability density is considered as density in the actual physical sense (in $\mathrm{kg/m^3}$) the expectation is the center of gravity. In the world of physics Steiner (1840) proved a result for the moment of inertia which translates – exploiting this analogy – to Steiner's identity (Brachinger 1999) which is widely known and used in statistics. Let for this purpose e_t be a series of realizations of a random variable, $\bar{e}$ their average, and d an arbitrary constant; it then follows that

$$\sum_{t=1}^{T}(e_t - d)^2 = \sum_{t=1}^{T}(e_t - \bar{e})^2 + T(\bar{e} - d)^2. \tag{3.1}$$

With slight abuse of notation, let e_t be the (out-of-sample) forecast error of a particular method at time t and denote the evaluation period by $t = 1, \ldots, T$. Setting $d = 0$, equation (3.1) becomes

$$\text{MSPE} = \frac{1}{T}\sum_{t=1}^{T} e_t^2 = \frac{1}{T}\sum_{t=1}^{T}(e_t - \bar{e})^2 + \bar{e}^2 = V + B^2. \tag{3.2}$$

The first term, V, represents the variance of the *forecast errors* while the second term, B^2, represents the squared bias of the *forecasts* generated from a particular method.

Fig. 3.3 shows the MSPE decomposition from equation (3.2) for selected methods, across all variables and forecast horizons. For the three interest rate series and HOUSING, the mean survey forecast is severely biased which accounts for a nontrivial share of the MSPE (often around 20% or more). By comparison with Tab. 3.4, these series are the ones where *TreeBagg* provides the clearest MSPE improvements over *EW*. These improvements can be explained by substantial bias reductions; observe from Fig. 3.3 that the bias component of

[14] Except for the variable TBOND which is only available since 1992Q1 and thus has a smaller training window sample for the first evaluation points.

TreeBagg for the interest rate series and HOUSING is negligible. In contrast, the variance of the forecast errors is not generally smaller for *TreeBagg* than for *EW*.

These findings show that when mean survey forecasts are severely biased, their performance can be clearly improved by *TreeBagg*. Interestingly, the converse also holds true: When survey forecasts are *not* severely biased, *TreeBagg* cannot improve upon *EW*. This is the case for CPI and NGDP. Fig. 3.3 shows that for these series, the MSPE of *EW* contains virtually no bias component; at the same time, Tab. 3.4 shows that *EW* tends to outperform *TreeBagg*. This result is driven by the fact that the forecast errors generated by *TreeBagg* have larger variance than those generated by *EW* (again see Fig. 3.3).

It is natural to compare the micro-level bias correction provided by *TreeBagg* to the aggregate-level bias correction provided by *CapTim*. For TBILL and BOND, Fig. 3.3 shows that *CapTim* performs the same role as *TreeBagg*: It corrects the bias in *EW*, at the cost of an acceptable variance increase. *CapTim* does not work well for TBOND and HOUSING, however. For these two series, *CapTim* either fails to correct the biases in *EW*, or it does so at the cost of a large variance increase.

Unlike the cases of *TreeBagg* and *CapTim*, the relative performance of *SlctBagg* vis-à-vis *EW* appears unsystematic across series and forecast horizons. As expected (see Section 3.2.4), Fig. 3.3 reveals that the bias component of *SlctBagg* is generally very similar in size to that of *EW*. Furthermore, the forecast error variance is not generally smaller (or larger) for *SlctBagg* than for *EW*. We view these findings as evidence that differences in the individual predictive ability of forecasters are too small and/or unsystematic to be exploited by schemes like *SlctBagg*. Together with the above comparison between *TreeBagg* and *CapTim*, this suggests that the gains of *TreeBagg* over *EW* are due to bias corrections, and *not* due to the trees' use of individual-level information.

Figure 3.3: MSPE decomposition

Decompositions of the MSPEs of selected forecasting methods as in Equation (3.2), for all series and forecast horizons. Percentage shares of the variance component (V/MSPE $\cdot$ 100) in (3.2)) are given inside the bars.

3.4.3 Results for the Pre-Crisis Sample

Our evaluation sample contains the recent financial crisis which is clearly reflected in the series we analyze. Under squared error loss and a small evaluation sample typical of empirical macroeconomics, it is important to analyze the impact of the crisis on our results. In Tab. C.3, we therefore re-evaluate our findings for a smaller evaluation sample ending in 2007Q3 ("pre-crisis sample").

First, we find that the benefits of using *TreeBagg* are smaller in the pre-crisis sample than in the complete sample: While *TreeBagg* outperforms *EW* in 23/35 comparisons in the complete sample, this is true for only 13/35 comparisons in the pre-crisis sample. Second, our finding that bagging improves the performance of *Tree* and *Slct* is equally valid for the pre-crisis sample (this is reflected in Tab. C.1 and Tab. C.2 in the Appendix). Third, for HOUSING and TBILL the performance of *AR* (relative to *EW*) is worse in the pre-crisis sample than in the complete sample; it is largely unchanged for the other variables.

Taken together, we find some evidence that equally weighted SPF forecasts perform better in the pre-crisis than in the complete sample; this is true both relative to *TreeBagg* and relative to *AR*. Conversely, the relative performance of *EW* during the crisis periods is quite poor. This casts some doubt on the usefulness of the "raw" *EW* survey forecasts in turbulent times and emphasizes the potential value of robust bias correction methods like *TreeBagg*.

3.5 Conclusion

This paper analyzes the combination of macroeconomic expert forecasts. The main scheme we propose, (bagged) regression trees, can estimate flexible conditional mean functions based on a large number of predictors without suffering from the curse of dimensionality. In the present context, this enables trees to correct potential biases of survey forecasts at the level of individual participants. This stands in contrast to simple methods like the mean or median which are summary statistics from the (anonymous) set of available forecasts.

Using data from the Survey of Professional Forecasters and an evaluation sample between 2001 and 2010, we find that trees achieve lower Mean Squared Prediction Errors than the mean survey forecast for the majority of time series and forecast horizons we consider. We show that these improvements occur whenever the survey forecasts are severely biased. Comparisons with other forecast combination methods suggest that these bias corrections – rather than the use of individual-level information – explain the improvements of trees over the mean forecast.

Finally, the applicability of trees as a forecast combination method is by no means limited to the setting we consider in this paper: First, other expert surveys like the Livingston survey or the European Survey of Professional Forecasters share the salient features of the Survey of Professional Forecasters. Second, while trees *can* deal with high-dimensional structures and missingness, there is no reason why they should not work equally well (or better) in less ambitious data settings, such as the combination of forecasts from econometric models.

Bibliography

Akaike, H. (1970), 'Statistical Predictor Identification', *Annals of the Institute of Statistical Mathematics* **22**(1), 203–217.

Andersen, T., Dobrev, D. & Schaumburg, E. (2010), Jump-Robust Volatility Estimation using Nearest Neighbor Truncation, Working Paper, National Bureau of Economic Research.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Ebens, H. (2001), 'The distribution of realized stock return volatility', *Journal of Financial Economics* **61**(1), 43–76.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2000), Exchange rate returns standardized by realized volatility are (nearly) gaussian, Working Paper, National Bureau of Economic Research.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2001), 'The Distribution of Realized Exchange Rate Volatility', *Journal of the American Statistical Association* **96**(453), 42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2003), 'Modeling and Forecasting Realized Volatility', *Econometrica* **71**(2), 579–625.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Vega, C. (2007), 'Real-time price discovery in global stock, bond and foreign exchange markets', *Journal of International Economics* **73**(2), 251–277.

Ang, A., Bekaert, G. & Wei, M. (2007), 'Do macro variables, asset markets, or surveys forecast inflation better?', *Journal of Monetary Economics* **54**(4), 1163–1212.

Areal, N. M. P. C. & Taylor, S. J. (2002), 'The realized volatility of FTSE-100 futures prices', *Journal of Futures Markets* **22**(7), 627–648.

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

Asai, M., McAleer, M. & Medeiros, M. C. (2012), 'Asymmetry and Long Memory in Volatility Modeling', *Journal of Financial Econometrics* **10**(3), 495–512.

Audrino, F. & Bühlmann, P. (2001), 'Tree-Structured Generalized Autoregressive Conditional Heteroscedastic Models', *Journal of the Royal Statistical Society. Series B, (Statistical Methodology)* **63**(4), 727–744.

Audrino, F. & Colangelo, D. (2010), 'Semi-parametric forecasts of the implied volatility surface using regression trees', *Statistics and Computing* **20**(4), 421–434.

Audrino, F. & Knaus, S. (2012), Lassoing the HAR model: A Model Selection Perspective on Realized Volatility Dynamics, Working Paper, University of St. Gallen, School of Economics and Political Science. Economics Working Paper Series.

Audrino, F. & Medeiros, M. C. (2011), 'Modeling and forecasting short-term interest rates: The benefits of smooth regimes, macroeconomic variables, and bagging', *Journal of Applied Econometrics* **26**(6), 999–1022.

Baillie, R. T. & Bollerslev, T. (1989), 'The Message in Daily Exchange Rates: A Conditional-Variance Tale', *Journal of Business & Economic Statistics* **7**(3), 60–68.

Barndorff-Nielsen, O. E. & Shephard, N. (2002*a*), 'Estimating Quadratic Variation Using Realized Variance', *Journal of Applied Econometrics* **17**(5), 457–477.

Barndorff-Nielsen, O. & Shephard, N. (2002*b*), 'Econometric analysis of realized volatility and its use in estimating stochastic volatility models', *Journal of the Royal Statistical Society. Series B, (Statistical Methodology)* **17**(5), 253–280.

Barndorff-Nielsen, O. & Shephard, N. (2004), 'Power and bipower variation with stochastic volatility and jumps', *Journal of Financial Econometrics* **2**(1), 1–37.

Barndorff-Nielsen, O., Shephard, N. & Winkel, M. (2006), 'Limit theorems for multipower variation in the presence of jumps', *Stochastic processes and their applications* **116**(5), 796–806.

Bates, J. M. & Granger, C. W. J. (1969), 'The Combination of Forecasts', *Operational Research Quarterly* **20**(4), 451–468.

Bauwens, L., Hafner, C. M. & Laurent, S. (2012), *Handbook of volatility models and their applications*, Vol. 3, Wiley.

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

Bollerslev, T. (1986), 'Generalized Autoregressive Conditional Heteroskedasticity', *Journal of Econometrics* **31**(3), 307–327.

Bollerslev, T. & Wright, J. H. (2001), 'High-Frequency Data, Frequency Domain Inference, and Volatility Forecasting', *Review of Economics and Statistics* **83**(4), 596–602.

Box, G. & Cox, D. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society. Series B, (Statistical Methodology)* **26**(2), 211–252.

Brachinger, H. (1999), Steiner Identity, *in* S. Kotz, C. B. Read & D. L. Banks, eds, 'Encyclopedia of Statistical Sciences (Update, 3rd Volume)', Wiley & Sons, New York.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and Regression Trees*, Chapman & Hall/CRC.

Brockwell, P. & Davis, R. (1986), *Time Series: Theory and Methods*, Springer-Verlag New York, Inc., New York.

Bühlmann, P. (2004), Bagging, Boosting and Ensemble Methods, *in* J. Gentle, W. Härdle & Y. Mori, eds, 'Handbook of Computational Statistics', Springer, New York, pp. 877–907.

Bühlmann, P. & Van De Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer-Verlag New York Inc.

Bühlmann, P. & Yu, B. (2002), 'Analyzing Bagging', *Annals of Statistics* **30**(4), 927–961.

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

Chatterjee, A. & Lahiri, S. N. (2011), 'Bootstrapping lasso estimators', *Journal of the American Statistical Association* **106**(494), 608–625.

Chen, X. & Ghysels, E. (2011), 'NewsGood or Badand Its Impact on Volatility Predictions over Multiple Horizons', *Review of Financial Studies* **24**(1), 46–81.

Chen, Y., Härdle, W. & Pigorsch, U. (2010), 'Localized Realized Volatility Modeling', *Journal of the American Statistical Association* **105**(492), 1376–1393.

Christoffersen, P. F. (1998), 'Evaluating Interval Forecasts', *International Economic Review* **39**(4), 841–862.

Clemen, R. (1989), 'Combining forecasts: A review and annotated bibliography', *International Journal of Forecasting* **5**(4), 559–583.

Corsi, F. (2009), 'A Simple Approximate Long-Memory Model of Realized Volatility', *Journal of Financial Econometrics* **7**(2), 174–196.

Corsi, F., Audrino, F. & Renò, R. (2012), HAR Modelling for Realized Volatility Forecasting, *in* 'Volatility Models and Their Applications', Wiley.

Corsi, F. & Renò, R. (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.

Craioveanu, M. & Hillebrand, E. (2010), Why it is OK to use the HAR-RV (1,5,21) model, Working Paper, Louisiana State University.

Croushore, D. (2006), Forecasting with Real-Time Macroeconomic Data, *in* C. G. G. Elliott & A. Timmermann, eds, 'Handbook of Economic Forecasting', Elsevier, pp. 961–982.

Diebold, F. X. & Mariano, R. S. (1995), 'Comparing Predictive Accuracy', *Journal of Business & Economic Statistics* **13**(3), 253–263.

Diebold, F. X. & Yilmaz, K. (2009), 'Measuring Financial Asset Return and Volatility Spillovers, with Application to Global Equity Markets', *Economic Journal* **119**(534), 158–171.

Diebold, F. X. & Yilmaz, K. (2012), 'Better to give than to receive: Predictive directional measurement of volatility spillovers', *International Journal of Forecasting* **28**(1), 57–66.

Dimpfl, T. & Jung, R. C. (2012), 'Financial market spillovers around the globe', *Applied Financial Economics* **22**(1), 45–57.

Engle, R. F. (1982), 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation', *Econometrica* **50**(4), 987–1007.

Engle, R. F., Gallo, G. M. & Velucchi, M. (2012), 'Volatility spillovers in East Asian financial markets: a MEM-based approach', *Review of Economics and Statistics* **94**(1), 222–223.

Engle, R. F. & Ng, V. K. (1993), 'Measuring and Testing the Impact of News on Volatility', *Journal of Finance* **48**(5), 1749–1778.

Faust, J. & Wright, J. (2011), Forecasting Inflation, Working Paper, John Hopkins University. Working Paper.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.

FTSE Group (2013), *FTSE 100 Index*. http://www.ftse.com/Analytics/FactSheets/Home/DownloadSingleIssue?issueName=UKX, accessed 2013-04-16.

Gauß, C. F. (1887), *Abhandlungen zur Methode der kleinsten Quadrate*, P. Stankiewicz.

Gey, S. & Nedelec, E. (2005), 'Model selection for CART regression trees', *IEEE Transactions on Information Theory* **51**(2), 658–670.

Giot, P. & Laurent, S. (2004), 'Modelling daily Value-at-Risk using realized volatility and ARCH type models', *Journal of Empirical Finance* **11**(3), 379–398.

Gębka, B. & Serwa, D. (2007), 'Intra- and inter-regional spillovers between emerging capital markets around the world', *Research in International Business and Finance* **21**(2), 203–221.

Golosnoy, V., Gribisch, B. & Liesenfeld, R. (2012), Intra-daily volatility spillovers between the US and German stock markets, Working Paper, Christian-Albrechts-Universität zu Kiel.

Gonçalves, S. & Meddahi, N. (2011), 'Box-Cox transforms for realized volatility', *Journal of Econometrics* **160**(1), 129–144.

Granger, C. W. J. (1969), 'Investigating Causal Relations by Econometric Models and Cross-spectral Methods', *Econometrica* **37**(3), 424–438.

Guerrero, V. M. (1993), 'Time-series analysis supported by power transformations', *Journal of Forecasting* **12**(1), 37–48.

Gupta, S. & Wilton, P. C. (1988), 'Combination of Economic Forecasts: An Odds-Matrix Approach', *Journal of Business & Economic Statistics* **6**(3), 373–379.

Halbleib, R. & Pohlmeier, W. (2012), 'Improving the value at risk forecasts: Theory and evidence from the financial crisis', *Journal of Economic Dynamics and Control* **36**(8), 1212–1228.

Hamao, Y., Masulis, R. & Ng, V. (1990), 'Correlations in Price Changes and Volatility across International Stock Markets', *Review of Financial Studies* **3**(2), 281–307.

Hamilton, J. (1994), *Time Series Analysis*, Princeton University Press.

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

Hansen, P. & Lunde, A. (2005), 'A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?', *Journal of Applied Econometrics* **20**(7), 873–889.

Hansen, P., Lunde, A. & Nason, J. (2011), 'The Model Confidence Set', *Econometrica* **79**(2), 453–497.

Hart, J. D. & Wehrly, T. E. (1986), 'Kernel Regression Estimation Using Repeated Measurements Data', *Journal of the American Statistical Association* **81**(396), 1080–1088.

Hastie, T. & Efron, B. (2011), *lars: Least Angle Regression, Lasso and Forward Stagewise*.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn, Springer.

Hillebrand, E. & Medeiros, M. C. (2010), 'The Benefits of Bagging for Forecast Models of Realized Volatility', *Econometric Reviews* **29**(5-6), 571–593.

Hilliard, J. E. (1979), 'The Relationship Between Equity Indices on World Exchanges', *Journal of Finance* **34**(1), 103–114.

Hintze, J. & Nelson, R. (1998), 'Violin Plots: A Box Plot-density Trace Synergism', *American Statistician* **52**(2), 181–184.

Hong-Zhi, A., Zhao-Guo, C. & Hannan, E. J. (1982), 'Autocorrelation, Autoregression and Autoregressive Approximation', *Annals of Statistics* **10**(3), 926–936.

Inoue, A. & Kilian, L. (2008), 'How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation', *Journal of the American Statistical Association* **103**(482), 511–522.

Ishida, I. & Toshiaki, W. (2009), Modeling and Forecasting the Volatility of the Nikkei 225 realized Volatility using the ARFIMA-GARCH model, Working Paper, University of Tokyo. CIRJE F-Series.

Jarque, C. M. & Bera, A. K. (1980), 'Efficient tests for normality, homoscedasticity and serial independence of regression residuals', *Economics Letters* **6**(3), 255–259.

Jiang, G. J., Konstantinidi, E. & Skiadopoulos, G. (2012), 'Volatility spillovers and the effect of news announcements', *Journal of Banking & Finance* **36**(8), 2260–2273.

King, M. A. & Wadhwani, S. (1990), 'Transmission of Volatility between Stock Markets', *Review of Financial Studies* **3**(1), 5–33.

Kock, A. B. (2012), On the Oracle Property of the Adaptive LASSO in Stationary and Nonstationary Autoregressions, Working Paper, Aarhus University. CREATES Research Paper 2012-05.

Kock, A. B. & Callot, L. A. (2012*a*), Oracle Efficient Estimation and Forecasting with the Adaptive LASSO and the Adaptive Group LASSO in Vector Autoregressions, Working Paper, Aarhus University. CREATES Research Paper 2012-38.

Kock, A. B. & Callot, L. A. (2012*b*), Oracle Inequalities for High Dimensional Vector Autoregressions, Working Paper, Aarhus University. CREATES Research Paper 2012-16.

Kuester, K., Mittnik, S. & Paolella, M. S. (2006), 'Value-at-risk prediction: A comparison of alternative strategies', *Journal of Financial Econometrics* **4**(1), 53–89.

Kupiec, P. (1995), 'Techniques for Verifying the Accuracy of Risk Measurement Models', *Journal of Derivatives* **3**(2), 73–84.

Leeb, H. & Pötscher, B. (2005), 'Model selection and inference: Facts and fiction', *Econometric Theory* **21**(1), 21–59.

Legendre, A. M. (1805), *Nouvelles méthodes pour la détermination des orbites des comètes*, F. Didot.

Lewis, D. (1991), *Matrix theory*, World Scientific Publishing Company Inc.

Marcellino, M., Stock, J. H. & Watson, M. W. (2003), 'Macroeconomic Forecasting in the Euro area: Country Specific versus Area-Wide Information', *European Economic Review* **47**(1), 1–18.

Marcellino, M., Stock, J. H. & Watson, M. W. (2006), 'A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series', *Journal of Econometrics* **135**(1-2), 499–526.

Martens, M., van Dijk, D. & de Pooter, M. (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.

McAleer, M. & Medeiros, M. C. (2008), 'Realized Volatility: A Review.', *Econometric Reviews* **27**(1-3), 10–45.

Mincer, J. & Zarnowitz, V. (1969), The Evaluation of Economic Forecasts, *in* J. Mincer, ed., 'Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance', National Bureau of Economic Research, chapter The Evaluation of Economic Forecasts, pp. 1–46.

Morgan, J. N. & Sonquist, J. A. (1963), 'Problems in the Analysis of Survey Data, and a Proposal', *Journal of the American Statistical Association* **58**(302), 415–434.

Nardi, Y. & Rinaldo, A. (2011), 'Autoregressive process modeling via the Lasso procedure', *Journal of Multivariate Analysis* **102**(3), 528–549.

Newey, W. K. & West, K. D. (1987), 'A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix', *Econometrica* **55**(3), 703–708.

Newey, W. K. & West, K. D. (1994), 'Automatic Lag Selection in Covariance Matrix Estimation', *Review of Economic Studies* **61**(4), 631–653.

Nikkei (2013), *Nikkei Constituents*. http://e.nikkei.com/e/app/fr/market/constituents.aspx; accessed 2013-04-16.

Nikkinen, J., Omran, M., Sahlström, P. & Äijö, J. (2006), 'Global stock market reactions to scheduled US macroeconomic news announcements', *Global Finance Journal* **17**(1), 92–104.

Pankratz, A. & Dudley, U. (1987), 'Forecasts of Power-transformed Series', *Journal of Forecasting* **6**(4), 239–239.

Park, H. & Sakaori, F. (2013), 'Lag weighted lasso for time series model', *Computational Statistics* **28**(2), 493–504.

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

Patton, A. J. & Timmermann, A. (2010), 'Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion', *Journal of Monetary Economics* **57**(7), 803–820.

Pesaran, M. H. & Weale, M. (2006), Survey Expectations, *in* C. G. G. Elliott & A. Timmermann, eds, 'Handbook of Economic Forecasting', Elsevier, pp. 715–776.

Politis, D. N. & Romano, J. P. (1994), 'The Stationary Bootstrap', *Journal of the American Statistical Association* **89**(428), 1303–1313.

Poncela, P., Rodríguez, J., Sánchez-Mangas, R. & Senra, E. (2011), 'Forecast combination through dimension reduction techniques', *International Journal of Forecasting* **27**(2), 224–237.

Poors, S. . (2013), *S& P Dow Jones Indices - Fact Sheet*. http://www.spindices.com/documents/factsheets/fs-sp-500-ltr.pdf, accessed 2013-04-16.

Proietti, T. & Lütkepohl, H. (2013), 'Does the Box–Cox transformation help in forecasting macroeconomic time series?', *International Journal of Forecasting* **29**(1), 88–99.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rapach, D. & Strauss, J. (2010), 'Bagging or combining (or both)? An analysis based on forecasting US employment growth', *Econometric Reviews* **29**(5), 511–533.

Rea, W., Reale, M., Cappelli, C. & Brown, J. (2010), 'Identification of Changes in Mean with Regression Trees: An Application to Market Research', *Econometric Reviews* **29**(5), 754–777.

Ripley, B. (2008), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

Rossi, A. & Timmermann, A. (2010), What is the Shape of the Risk-Return Relation?, Working Paper, University of California San Diego. Working Paper.

Savva, C. S., Osborn, D. R. & Gill, L. (2009), 'Spillovers and correlations between US and major European stock markets: the role of the euro', *Applied Financial Economics* **19**(19), 1595–1604.

Scharth, M. & Medeiros, M. C. (2009), 'Asymmetric effects and long memory in the volatility of Dow Jones stocks', *International Journal of Forecasting* **25**(2), 304–327.

Shapiro, S. S. & Wilk, M. B. (1965), 'An analysis of variance test for normality (complete samples)', *Biometrika* **52**(3-4), 591–611.

Smith, J. & Wallis, K. (2009), 'A Simple Explanation of the Forecast Combination Puzzle', *Oxford Bulletin of Economics and Statistics* **71**(3), 331–355.

SPF (2010), Survey of Professional Forecasters - Documentation. Available at http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/spf-documentation.pdf. Dated September 8, 2010.

Stark, T. (2010), Realistic Evaluation of Real-Time Forecasts in the Survey of Professional Forecasters, Working Paper, Federal Reserve Bank of Philadelphia.

Steiner, J. (1840), 'Von dem Krümmungs-Schwerpuncte Ebener Curven', *Journal für die Reine und Angewandte Mathematik* **21**(1), 33–63.

Tabarrok, A. (2013), 'Twitter, 10:32 am, 17 apr 13, cet'. accessed 2013-04-28. `https://twitter.com/ATabarrok/status/324576147348267008`

Therneau, T., Atkinson, B. & Ripley, B. (2010), *rpart: Recursive Partitioning*. R package version 3.1-46.

Thomakos, D. D. & Wang, T. (2003), 'Realized volatility in the futures markets', *Journal of Empirical Finance* **10**(3), 321–353.

Tibshirani, R. (1996), 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society. Series B, (Statistical Methodology)* **58**(1), 267–288.

Timmermann, A. (2006), Forecast Combinations, *in* G. Elliott, C. Granger & A. Timmermann, eds, 'Handbook of Economic Forecasting', Elsevier, Amsterdam, pp. 135–196.

Zhang, L., Mykland, P. & Aït-Sahalia, Y. (2005), 'A Tale of Two Time Scales', *Journal of the American Statistical Association* **100**(472), 1394–1411.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of Lasso', *Journal of Machine Learning Research* **7**(2), 2541–2563.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

Appendix A

Lassoing the HAR model: A Model Selection Perspective on Realized Volatility Dynamics

A.1 Proof of Theorem 2

This proof is structured as follows. We first show in Lemma 1 that the irrepresantable condition is satisfied for the HAR model. Based on this we invoke a theorem of Zhao & Yu (2006) which relaxes the assumptions on the innovation term for the lasso to be model consistent. Finally we show that the HAR model satisfies the assumptions of the aforementioned theorem and we can thus expect the lasso to be model selection consistent without the assumption Gaussianity for the error term.

Lemma 1. *Under the assumption that HAR model is true, condition (ii) of Theorem 1 is satisfied.*

Lemma 1 states that if the true DGP indeed obeys the law of motion as specified by the HAR model one can apply the results of Nardi & Rinaldo (2011) who establish that the lasso is a valid model selection device under two assumption, namely, that (i) $\|\Gamma_{SS}^{-1}\|_\infty \leq C$ and (ii) $\|\Gamma_{S^cS}\Gamma_{SS}^{-1}\|_\infty < 1$. Γ denotes the autocovariance matrix, S is the true active set of predictors, S^c is the true non-active set of predictors. When embedding the HAR model in this specification we have that S consists of the lagged values up to order 22 and S^c is any other lagged values beyond 22. Since (i) holds trivially as by (1.1) none of variables is a linear combination of another, we only collect the proof of (ii) in the Lemma below.

Proof. The proof is split into two parts. First we show that the infinity norm of $\Gamma_{S^cS}\Gamma_{SS}^{-1}$ can be seen as the sum of the absolute values of the regression coefficients of the usual HAR estimates, second, we show that it is sufficient to consider one specific non-active regressor.

Moreover, consider the following equivalent notations:

$$\mathrm{Cov}(S^c, S)\,\mathrm{Var}(S)^{-1} = \mathrm{Cov}(S^c, S)\,\mathrm{Cov}(S, S)^{-1} = \Gamma_{S^cS}\Gamma_{SS}^{-1}.$$

To rule out any possible confusion we re-state the definition of the infinity norm of a matrix. If $\|\xi\|_\infty$ for $\xi \in \mathbb{R}^n$ is defined as $\|\xi\|_\infty = \max_{1\leq i\leq n} |\xi_i|$, then the corresponding matrix norm is given as

$$\|A\|_\infty := \max_{\|\xi\|_\infty = 1} \|A\xi\|_\infty$$

where it can be shown (Lewis 1991, Proposition 3.4.1) for $A = [a_{ij}]_{1\leq i\leq n, 1\leq j\leq m}$ that

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{m} |a_{ij}|.$$

In what follows we consider a row-vector $\xi = [\xi_1, \ldots, \xi_n]$ as $1 \times n$ matrix such that $\|\xi\|_\infty = \|\xi'\|_1$.

Throughout the proof we assume without loss of generality the HAR model to contain no intercept. Moreover, for the sake of notational simplicity we assume the AR process to be labeled as

$$x_t = \sum_{i=1}^{22} \phi_i x_{t-i} + \varepsilon_t. \tag{A.1}$$

Assume that $|S^c| = 1$ with $S^c = \{x_{t-23}\}$[1] and that the true model is in fact the HAR model, i.e. $|S| = 22$ with $S = \{x_{t-1}, x_{t-2}, \ldots, x_{t-22}\}$. In other words, the active set consists of the first 22 lagged values and the first non-active predictor is x_{t-23}. We then find that

$$\mathrm{Cov}(x_{t-23}, [x_{t-1}, x_{t-2}, \ldots, x_{t-22}])\mathrm{Var}([x_{t-1}, x_{t-2}, \ldots, x_{t-22}])^{-1} = [\tilde{\phi}_1, \ldots, \tilde{\phi}_{22}], \tag{A.2}$$

where $[\tilde{\phi}_1, \ldots, \tilde{\phi}_{22}]$ is the usual representation of regression coefficients of x_{t-23} on x_{t-1}, x_{t-2}, …, x_{t-22} (note that the previously introduced superscript "HAR" is omitted to alleviate notation).

Since we are only interested in the sum of the absolute values of these regression coefficients, i.e. $\|[\tilde{\phi}_1, \ldots, \tilde{\phi}_{22}]\|_\infty$, we may as well reorder the regressors since

$$\|[\tilde{\phi}_1, \ldots, \tilde{\phi}_{22}]\|_\infty = \|[\tilde{\phi}_{\sigma(1)}, \ldots, \tilde{\phi}_{\sigma(22)}]\|_\infty \tag{A.3}$$

is true for any permutation σ. With $\sigma(i) = 22 - i + 1$ we find that

$$\|[\tilde{\phi}_{\sigma(1)}, \ldots, \tilde{\phi}_{\sigma(22)}]\|_\infty = \|\mathrm{Cov}(x_{t-23}, [x_{t-22}, x_{t-21}, \ldots, x_{t-1}])\mathrm{Var}([x_{t-22}, x_{t-21}, \ldots, x_{t-1}])^{-1}\|_\infty$$

A closer look at the second term (exploiting covariance stationarity and thus, the fact that

[1]Observe that we slightly deviate from the notation used previously where $S \subset \mathbb{N}$; we use S and S^c to denote the corresponding lags variables rather than their indices.

the autocovariance is an even function, (see for instance Brockwell & Davis (1986)) shows that

$$\begin{aligned}
\mathrm{Cov}(x_{t-23}, [x_{t-22}, x_{t-21}, \ldots, x_{t-1}] &= [\mathrm{Cov}(x_{t-23}, x_{t-(23-i)})]_{1\leq i\leq 22} \\
&= [\mathrm{Cov}(x_t, x_{t-i})]_{1\leq i\leq 22} \\
&= \mathrm{Cov}(x_t, [x_{t-1}, x_{t-2}, \ldots, x_{t-22}])
\end{aligned}$$

and

$$\mathrm{Var}([x_{t-22}, x_{t-21}, \ldots, x_{t-1}]) = \mathrm{Var}([x_{t-1}, x_{t-2}, \ldots, x_{t-22}])$$

such that

$$\begin{aligned}
[\tilde{\phi}_{\sigma(1)}, \tilde{\phi}_{\sigma(2)}, \ldots, \tilde{\phi}_{\sigma(22)}] &= \mathrm{Cov}(x_{t-23}, [x_{t-22}, x_{t-21}, \ldots, x_{t-1}])\,\mathrm{Var}([x_{t-22}, x_{t-21} \ldots, x_{t-1}) \\
&= \mathrm{Cov}(x_t, [x_{t-1}, x_{t-2}, \ldots, x_{t-22}])\,\mathrm{Var}([x_{t-1}, x_{t-2}, \ldots, x_{t-22}])^{-1} \\
&= [\phi_1, \phi_2, \ldots, \phi_{22}] \qquad \text{(A.4)}
\end{aligned}$$

Combining (A.3) and (A.4) shows that (A.2) is indeed simply the sum of the absolute values of the coefficients of (A.1), i.e., we conclude for $S^c = \{x_{t-23}\}$ that we have

$$||\Gamma_{S^cS}\Gamma_{SS}^{-1}||_\infty = \beta^{(d)} + \beta^{(w)} + \beta^{(m)}. \qquad \text{(A.5)}$$

When extending the set of non-active predictors to $S^C = \{x_{t-(22+i)}\}_{1\leq i\leq k}$ one can verify[2] that

$$\begin{aligned}\mathrm{Cov}([x_{t-(22+1)},\ldots,x_{t-(22+k)}],[x_{t-1},x_{t-2},\ldots,x_{t-22}])\mathrm{Var}([x_{t-1},x_{t-2},\ldots,x_{t-22}])^{-1}\\ =\begin{bmatrix}\tilde{\phi}_1^{(1)} & \tilde{\phi}_2^{(1)} & \cdots & \tilde{\phi}_{22}^{(1)}\\ \vdots & \vdots & & \vdots\\ \tilde{\phi}_1^{(k)} & \tilde{\phi}_2^{(k)} & \cdots & \tilde{\phi}_{22}^{(k)}\end{bmatrix}.\end{aligned} \tag{A.6}$$

Hence,

$$\|\mathrm{Cov}(S^c,S)\mathrm{Var}(S)^{-1}\|_\infty = \max_{1\leq j\leq k}\sum_{i=1}^{22}|\tilde{\phi}_i^{(j)}|.$$

In a next step we show that $\sum_{i=1}^{22}|\phi_i^{(l)}| < \sum_{i=1}^{22}|\phi_i^{(k)}|$ for $l > k$ by induction. The conclusion then follows since it holds for $k = 1$, i.e. for $S^c = x_{t-23}$ which has already been proved in (A.5).

Given the argument which shows that reversing the order has no effect on the sum of the coefficients we present the argument in the usual AR(22) representation as given in (A.1) and thus drop the tilde, i.e.

$$x_{t+j} = \sum_{i=1}^{22}\phi_i^{(j)}x_{t+j-i} + \varepsilon_{t+j}.$$

[2]This can either be seen by establishing the usual AR(p) moment conditions or recalling the fact that the OLS estimates of an AR(p) process are consistent.Note that the consistency of the AR(p) estimates only gives results a.s. by asymptotic equivalence. However, basing the argument on theoretical moments and the fact that for appropriate random matrices X and Y we have $[\mathrm{Cov}(Y,X)\,\mathrm{Var}(X)^{-1}]' = \mathrm{Var}(X)^{-1}\,\mathrm{Cov}(X,Y)$ yields (A.6) directly.

Now, consider the induction basis for $j = 1 \to 2$:

$$
\begin{aligned}
x_{t+1} &= \sum_{i=1}^{22} \phi_i^{(1)} x_{t+1-i} + \epsilon_{t+1} \\
&= \phi_1^{(1)} \left(\sum_{i=1}^{22} \phi_i^{(1)} x_{t-i} + \epsilon_t \right) + \sum_{i=2}^{22} \phi_i^{(1)} x_{t+1-i} + \epsilon_{t+1} \\
&= \sum_{i=1}^{21} \left(\phi_1^{(1)} \phi_i^{(1)} + \phi_{i+1}^{(1)} \right) x_{t-i} + \phi_1^{(1)} \phi_{22}^{(1)} x_{t-22} + \tilde{\epsilon}_{t+1} \\
&= \sum_{i=1}^{22} \phi_i^{(2)} x_{t-i} + \tilde{\epsilon}_{t+1},
\end{aligned}
$$

where $\tilde{\epsilon}_{t+1} = \phi_1^{(1)} \epsilon_t + \epsilon_{t+1}$ and

$$
\phi_i^{(2)} = \phi_1^{(1)} \phi_i^{(1)} + \phi_{i+1}^{(1)} \text{ for } i = 1, \ldots, 21 \text{ and } \phi_{22}^{(2)} = \phi_1^{(1)} \phi_{22}^{(1)}. \tag{A.7}
$$

By the assumptions put forward in (1.1) we have that $\phi_i^{(2)} > 0 \ \forall i = 1, \ldots, 22$ and taking the difference of the sum of absolute values thus yields

$$
\sum_{i=1}^{22} |\phi_i^{(2)}| - \sum_{i=1}^{22} |\phi_i^{(1)}| = \phi_1^{(1)} \left(\sum_{i=1}^{22} \phi_i^{(1)} - 1 \right) = \phi_1^{(1)} \left(\beta^{(d)} + \beta^{(w)} + \beta^{(m)} - 1 \right).
$$

By (A.7) and (1.3) we have the induction basis $\phi_i^{(2)} > 0 \ \forall i = 1 \ldots 22$ and also we find by the fact[3] $\beta^{(d)} + \beta^{(w)} + \beta^{(m)} < 1$ that $\sum_{i=1}^{22} \phi_i^{(j-1)} < \sum_{i=1}^{22} \phi_i^{(j)}$.

[3]This follows directly from the causality assumption: Since all roots lie outside the unit circle and the $P(z)$, the characteristic polynomial, is continous on $\mathbb{R}$ it follows that $P(1) > 0$ and thus that $\beta^{(d)} + \beta^{(w)} + \beta^{(m)} < 1$.

Reapplying the same argument for the induction step $j \to j+1$ yields

$$\begin{aligned}
x_{t+j} &= \sum_{i=1}^{22} \phi_i^{(j)} x_{t+1-i} + \varepsilon_{t+j} \\
&= \phi_1^{(j)} \left(\sum_{i=1}^{22} \phi_i^{(1)} x_{t-i} + \varepsilon_t \right) + \sum_{i=2}^{22} \phi_i^{(j)} x_{t+1-i} + \varepsilon_{t+j} \\
&= \sum_{i=1}^{21} \left(\phi_1^{(j)} \phi_i^{(1)} + \phi_{i+1}^{(j)} \right) x_{t-i} + \phi_{22}^{(1)} \phi_1^{(j)} x_{t-22} + \tilde{\varepsilon}_{t+j} \\
&= \sum_{i=1}^{22} \phi_i^{(j+1)} x_{t-i} + \tilde{\varepsilon}_{t+j}
\end{aligned}$$

where again $\tilde{\varepsilon}_{t+j} = \phi_1^{(1)} \varepsilon_t + \varepsilon_{t+j}$ and $\phi_i^{(j+1)} = \phi_1^{(j)} \phi_i^{(1)} + \phi_{i+1}^{(j)}$ for $i = 1, \dots, 21$ and $\phi_{22}^{(j+1)} = \phi_1^{(1)} \phi_{22}^{(j)}$.

Taking the difference between the sum of $\phi_i^{(j+1)}$ and the sum of $\phi_i^{(j)}$ yields

$$\sum_{i=1}^{22} \phi_i^{(j+1)} - \sum_{i=1}^{22} \phi_i^{(j)} = \left(\sum_{i=1}^{22} \phi_i^{(1)} - 1 \right) \phi_1^{(j)}.$$

By the induction basis we have $\phi_i^{(j)} > 0 \ \forall i = 1, \dots, 22$ such that $\phi_i^{(j+1)} > 0 \ \forall i = 1, \dots, 22$ and thus

$$\sum_{i=1}^{22} |\phi_i^{(j+1)}| - \sum_{i=1}^{22} |\phi_i^{(j)}| < 0$$

such that the claim

$$\sum_{i=1}^{22} |\phi_i^{(j+1)}| < \sum_{i=1}^{22} |\phi_i^{(j)}|$$

follows. Summarizing we conclude that for the HAR model it holds that $\|\Gamma_{S^cS}\Gamma_{SS}\|_\infty \leq 1-\delta$ if $\beta^{(d)} + \beta^{(w)} + \beta^{(m)} \leq 1 - \delta$. □

Having proven the above we look at a theorem provided by Zhao & Yu (2006) which shows that the lasso is model selection consistent under some assumptions. Later we will prove that these assumptions hold if the HAR model is assumed to be true and we can thus safely relax the assumption of normally distributed errors if we are willing to accept a fixed S and S^c (as opposed to Nardi & Rinaldo's (2011) results where $p = |S|$ is allowed to grow as the sample size increases.

Theorem A (Zhao & Yu (2006)). *Under the assumptions of S and S^c fixed and*

(A1) $|\Gamma_{S^C S}\Gamma_{SS}^{-1}\,\mathrm{sgn}(\mathrm{supp}\,\phi^0)| \overset{a.s.}{<} \mathbf{1}$ *where* $\mathbf{1}$ *is a vector of ones and the inequality is understood componentwise*

(A2) $\Gamma^n_{(S,S^c),(S,S^c)} \xrightarrow{a.s.} \Gamma_{(S,S^c),(S,S^c)}$ *where* $\Gamma_{(S,S^c)}$ *is the autocovariance matrix and* $\Gamma^n_{(S,S^c)}$ *its sample analogon*

(A3) $\frac{1}{n}\max_{0\le i\le n-p}\sum_{j=1}^{p} x^2_{t-i-j} \xrightarrow{a.s.} 0$

the lasso is model selection consistent in the sense of Definition 1 if the innovation term has finite second moment and λ_n *is chosen such that* $\lambda_n/n \to 0$ *and* $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ *with* $0 \le c < 1$.

Proof of Theorem 2. We prove that the assumptions of Theorem A above are satisfied if one assumes the dynamics of the HAR model as put forward in (1.1) to hold as well as the existence of a finite fourth moment of the innovation term.

(A1) $|\Gamma_{S^C S}\Gamma_{SS}^{-1}\,\mathrm{sgn}(\mathrm{supp}\,\phi^0)| \overset{a.s.}{<} \mathbf{1}$ in (A1) of Theorem A holds since the argument in the proof of Lemma 1 can be made in terms of sample moments. Knowing that the least squares estimates converge a.s. to the true values (Brockwell & Davis 1986, Theorem 10.8.1) the conclusion follows since $|\Gamma_{S^C S}\Gamma_{SS}^{-1}\,\mathrm{sgn}(\mathrm{supp}\,\phi^0)| \overset{a.s.}{<} \mathbf{1}$ is weaker than $\|\Gamma_{S^c S}\Gamma_{SS}^{-1}\|_\infty \le 1-\delta$ as all components of $\mathrm{supp}\,\phi^0$ are greater than zero by (1.3).

(A2) Under the assumption of a finite fourth moment of the innovations we have by a result of Hong-Zhi, Zhao-Guo & Hannan (1982) the convergence almost surely. The positive definiteness follows from the fact that $\Gamma_{(S,S^c)}$ is positive semi-definite iff a variable is a linear combination of the others which is ruled out by the assumption of the HAR model as given in (1.3).[4]

(A3) Assuming that x_i is finite almost surely gives that $\frac{1}{n}\max_{0\le i\le n-p}\sum_{j=1}^{p} x_{t-i-j}^2$ is of class $\mathbf{o}_{a.s.}(n)$.

The condition on the innovation follows from Hölder's inequality since we have that $L^4 \subset L^2$ such that it suffices to require a finite fourth moment of the error term. □

Summarizing we have that the lasso should detect the HAR model if we assume a finite fourth moment.

A.2 Log-Transformed Volatilities

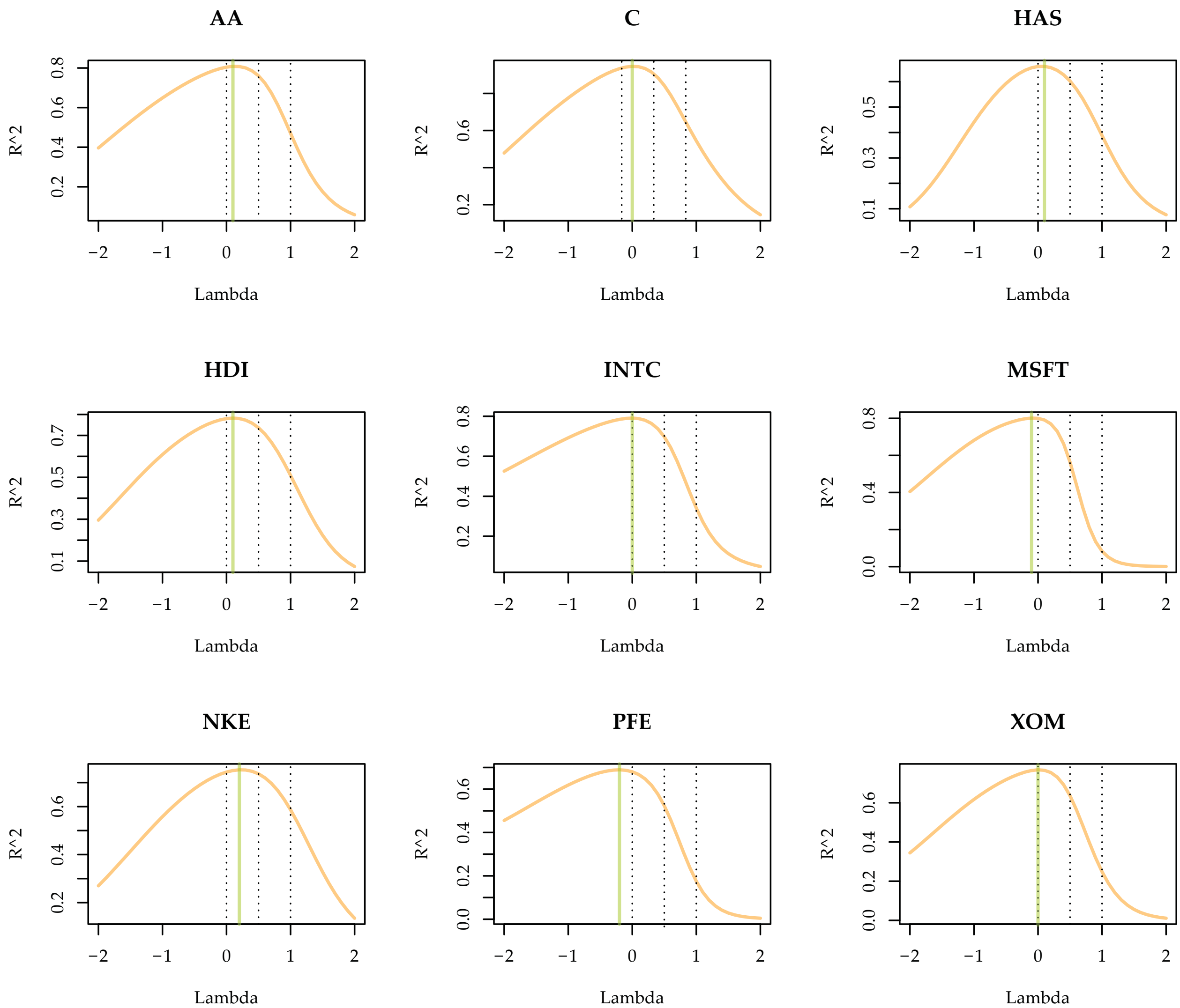
Although it is common to use the log-transform to model realized variance for reasons of positiveness, lower skewness and lower kurtosis, the case of the HAR model even allows for additional arguments to justify the use of log-transformed realized volatilities. These are not solely related to the realized volatility series as such (as for instance in Martens et al. (2009, Table 1)) but also to how realized volatility is modeled. Extending the approach of Box & Cox (1964) where only the dependent variable is transformed we employ the Box-Cox transform

$$f_\lambda(x) = x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{otherwise.} \end{cases}$$

[4] It is semi-definite since it is a covariance matrix.

to series of realized volatility. Consequently, the Box-Cox transform not only affects the dependent variable but also predictor variables in the HAR model. As in the original work of Box & Cox we then compute the (quasi-)likelihood for each λ. Since the (quasi-)likelihood is equivalent to the R^2 we report the R^2 for different values of λ in Fig. A.1.

Figure A.1: Coefficient of determination along power transform



R^2 for different values of λ for the HAR model estimated on $\mathrm{RV}_t^{(\lambda)}$ on the whole sample as described in Section 1.3.1. The green line indicates the maximal R^2 and the dotted lines indicate common transformations for realized volatilities ($\log \mathrm{RV}_t$ with $\lambda = 0$, $\sqrt{\mathrm{RV}_t}$ with $\lambda = 1/2$, and RV_t with $\lambda = 1$)

Clearly, following again Box & Cox and choosing a rational λ it follows that $\lambda = 0$

is a sensitive choice and thus justifies the use of log-transformed volatilities. A further argument for using $\lambda = 0$ may be found in the fact that for the case of $\lambda = 0$ we can construct unbiased estimates (under the assumption of normality of the log-transformed realized volatilities) explicitly without resorting to the median (Pankratz & Dudley 1987, Proietti & Lütkepohl 2013).

A.3 Robustness

This section shows the key results in graphical form as presented in the main paper if the realized volatility is estimated by Andersen et al.'s (2010) MedRV estimator instead of Zhang et al.'s (2005) two-time-scale estimator. MedRV is not only computationally attractive but also robust to zero returns and outliers induced by jumps. Figures A.2 to A.4 and Tab. A.1 are found below and are otherwise identical to the corresponding figures in the main text. There are marginal differences, but, the conclusions made in the main text remain valid such that we abstain from further discussion of these results.

Figure A.2: Autocorrelation function for $\log \mathrm{RV}_t$ series using MedRV.

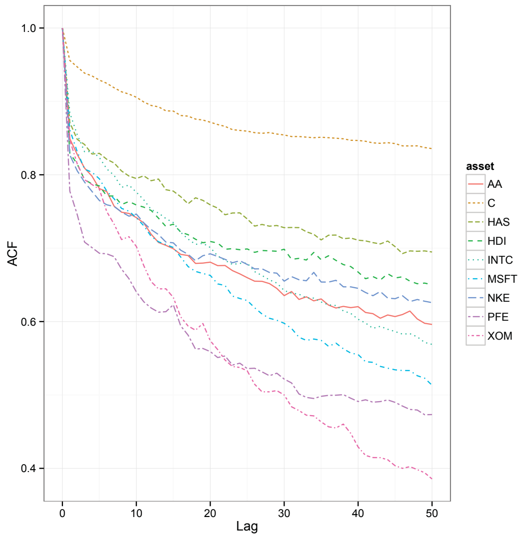


Figure A.2 (a)

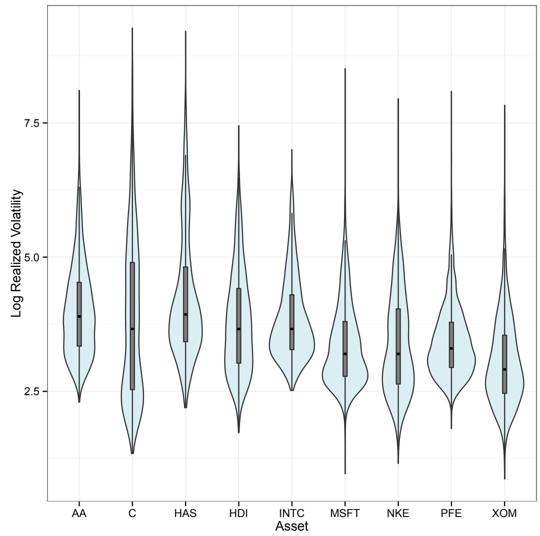


Figure A.2 (b)

Panel (a) shows the autocorrelation function for the 9 $\log \mathrm{RV}_t$ series. Panel (b) shows a violin plot (Hintze & Nelson 1998) of the unconditional $\log \mathrm{RV}_t$. Both use the MedRV estimator.

Figure A.3: HAR versus lasso coefficients with all predictors using MedRV estimator

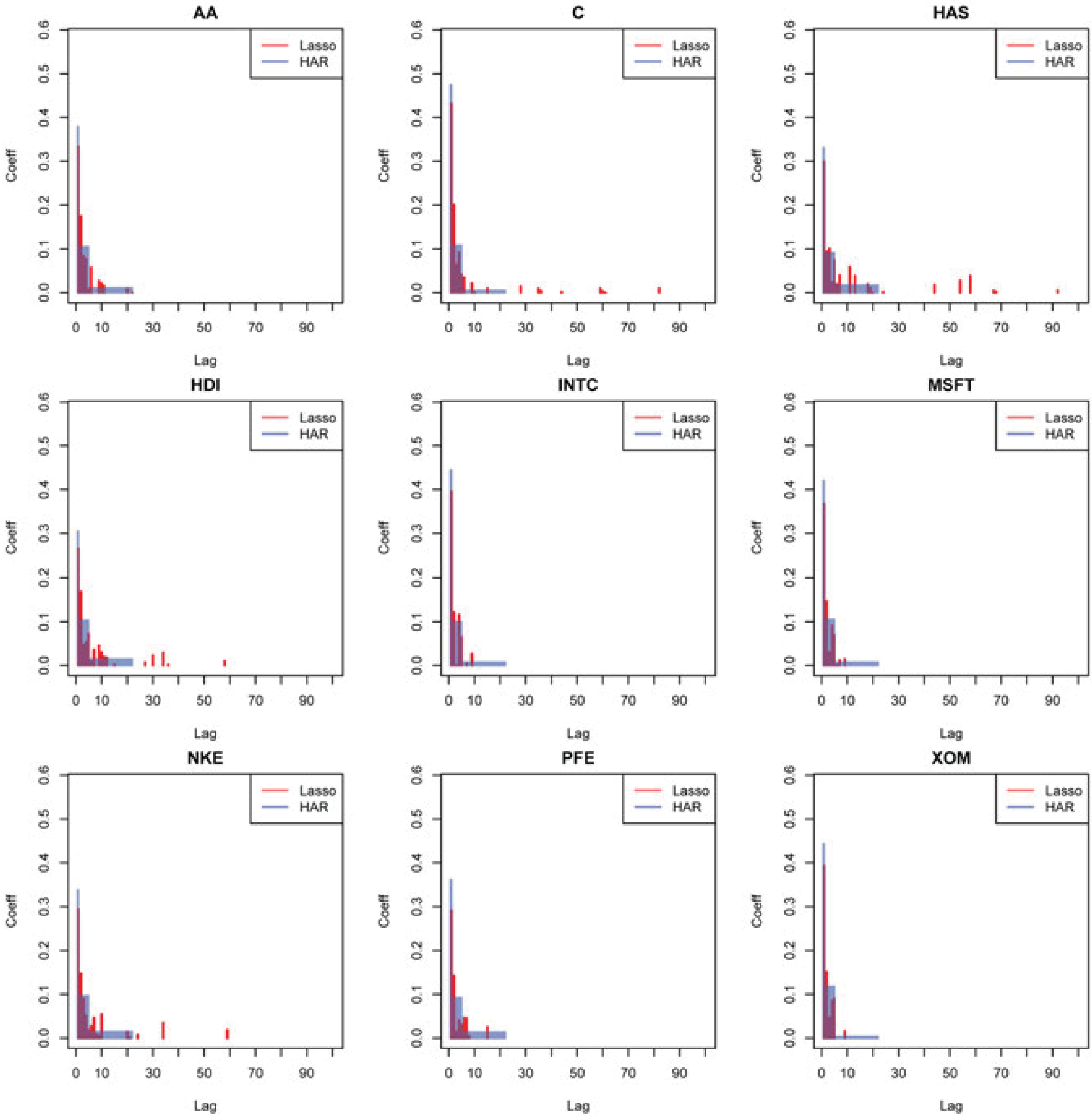


Figure A.4: Stability of Lasso selected Regressors for all assets using MedRV estimator

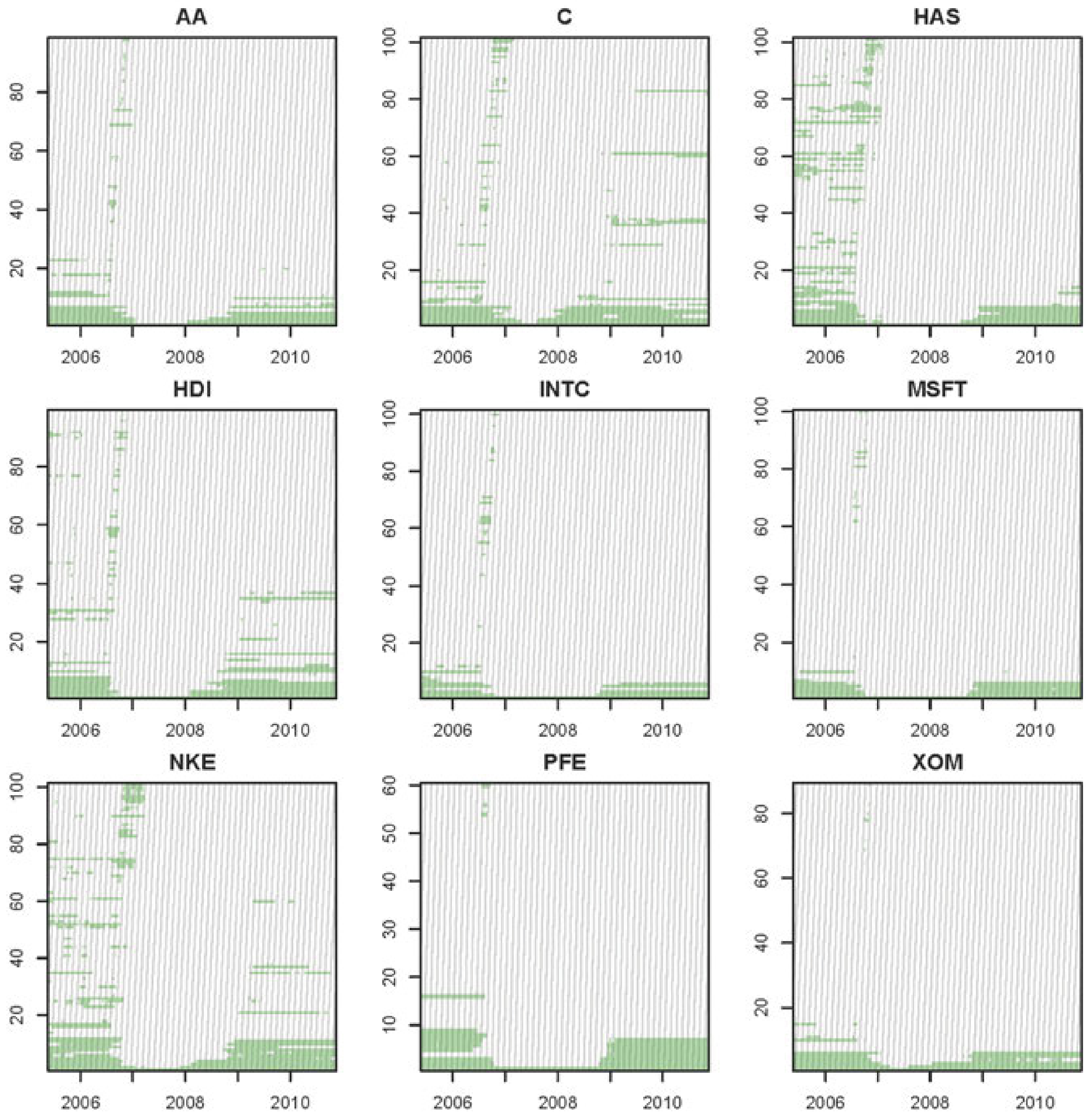


Table A.1: Diebold-Mariano tests of equal predictive ability

			AA	C	HAS	HDI	INTC	MSFT	NKE	PFE	XOM
1,000	Total	Mean Diff.	0.000	0.000	**-0.005**	-0.001	-0.001	**-0.002**	-0.004	**-0.003**	0.000
	(n=1'383)	*p*-value	*1.00*	*0.86*	*0.03*	*0.66*	*0.52*	*0.05*	*0.25*	*0.01*	*0.91*
	PreCrisis	Mean Diff.	-0.001	-0.003	-0.006	-0.002	0.000	-0.001	-0.005	-0.003	0.000
	(n=575)	*p*-value	*0.87*	*0.25*	*0.13*	*0.71*	*0.89*	*0.32*	*0.51*	*0.11*	*0.90*
	PostCrisis	Mean Diff.	0.000	0.003	-0.004	-0.001	-0.001	**-0.003**	-0.004	**-0.003**	0.000
	(n=808)	*p*-value	*0.88*	*0.48*	*0.13*	*0.81*	*0.40*	*0.08*	*0.29*	*0.06*	*0.96*
2,000	Total	Mean Diff.	0.001	-0.001	-0.003	0.002	0.000	0.001	0.000	0.001	0.003
	(n=383)	*p*-value	*0.70*	*0.21*	*0.39*	*0.57*	*0.82*	*0.61*	*0.95*	*0.56*	*0.31*
	PreCrisis	Mean Diff.	—	—	—	—	—	—	—	—	—
	(n=—)	*p*-value	—	—	—	—	—	—	—	—	—
	PostCrisis	Mean Diff.	0.001	-0.001	-0.003	0.002	0.000	0.001	0.000	0.001	0.003
	(n=383)	*p*-value	*0.70*	*0.21*	*0.39*	*0.57*	*0.82*	*0.61*	*0.95*	*0.56*	*0.31*

Difference in MSPE ($MSPE_{HAR} - MSPE_{lasso}$) are reported together with *p*-values from the Diebold-Mariano test (Diebold & Mariano 1995) with Newey-West adjusted standard errors (Newey & West 1987). The differences and *p*-values are reported for different training windows (1,000, 2,000) and before/after the financial crisis using the MedRV estimator.

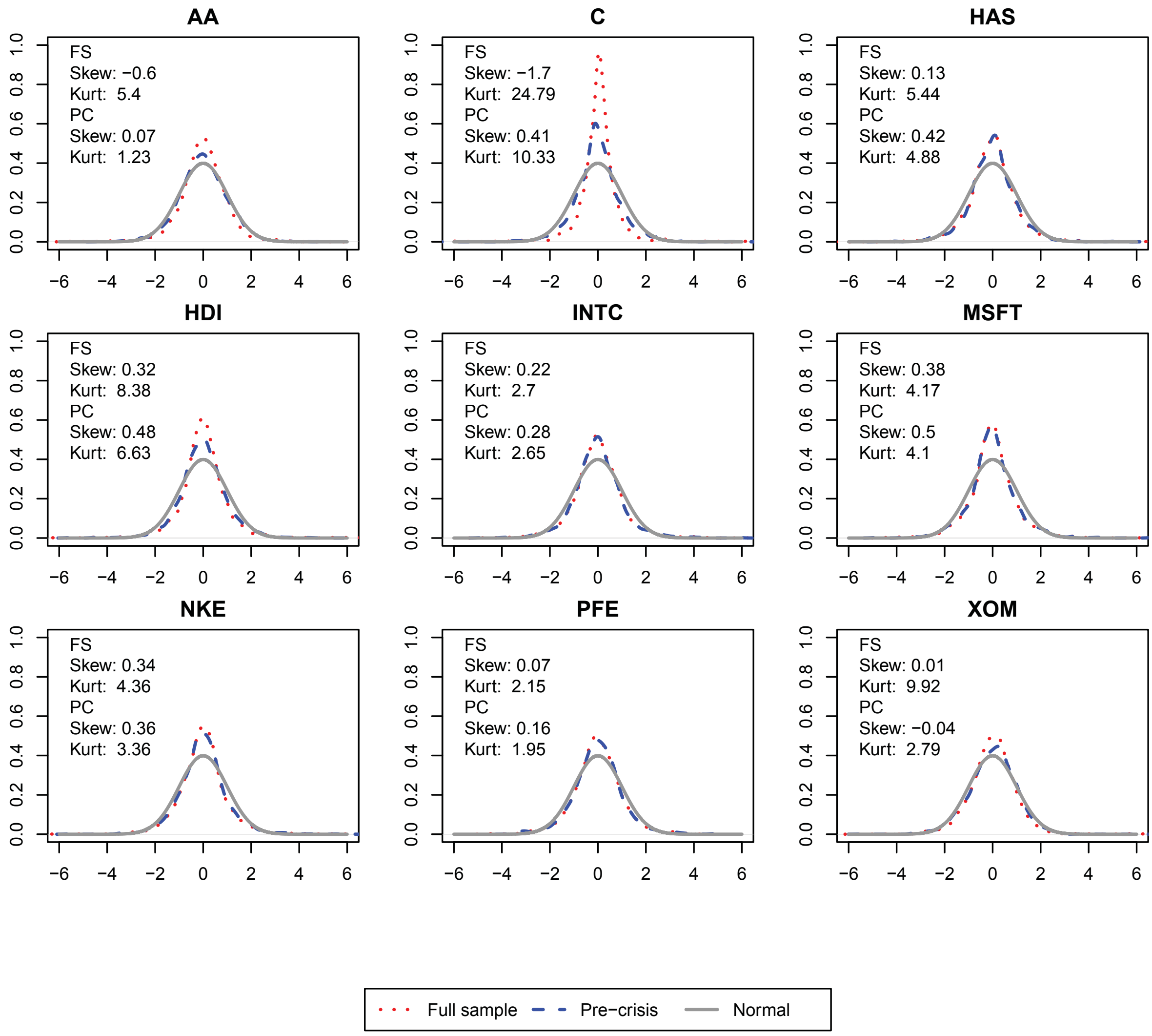
A.4 Risk Management Application

This section contains the actual violations of the value-at-risk visualized in Fig. 1.5 collected in Tab. A.2. Moreover, we have added summary statistics for the distribution of returns in Fig. A.5.

Table A.2: VaR violations (hit ratios)

				AA		C		HAS		HDI		INTC		MSFT		NKE		PFE		XOM	
	Model	α	Expected	Act.	p val	Act.	p val	Act.	p val	Act.	p val	Act.	p val	Act.	p val	Act.	p val	Act.	p val	Act.	p val
200	Emp	0.01	22	26	0.38	30	0.10	19	0.53	26	0.38	25	0.51	28	0.20	27	0.28	27	0.28	22	0.97
	HAR (Emp)	0.01	22	40	0.00	35	0.01	46	0.00	38	0.00	47	0.00	40	0.00	38	0.00	34	0.02	44	0.00
	HAR (Norm)	0.01	22	37	0.00	21	0.86	34	0.02	39	0.00	11	0.01	7	0.00	27	0.28	15	0.12	29	0.14
	Lasso (Emp)	0.01	22	46	0.00	35	0.01	48	0.00	38	0.00	51	0.00	47	0.00	43	0.00	37	0.00	47	0.00
	Lasso (Norm)	0.01	22	37	0.00	19	0.53	35	0.01	39	0.00	11	0.01	9	0.00	26	0.38	17	0.28	33	0.03
	Emp	0.025	55	56	0.85	68	0.08	56	0.85	61	0.39	60	0.46	55	0.95	62	0.32	54	0.94	55	0.95
	HAR (Emp)	0.025	55	72	0.02	71	0.03	76	0.01	72	0.02	81	0.00	79	0.00	78	0.00	66	0.13	80	0.00
	HAR (Norm)	0.025	55	66	0.13	42	0.07	59	0.55	75	0.01	36	0.01	27	0.00	62	0.32	32	0.00	59	0.55
	Lasso (Emp)	0.025	55	74	0.01	73	0.02	73	0.02	76	0.01	83	0.00	77	0.00	76	0.01	72	0.02	83	0.00
	Lasso (Norm)	0.025	55	68	0.08	41	0.05	56	0.85	82	0.00	36	0.01	23	0.00	65	0.17	34	0.00	59	0.55
400	Emp	0.01	20	28	0.08	40	0.00	20	0.97	31	0.02	21	0.79	23	0.49	24	0.36	23	0.49	21	0.79
	HAR (Emp)	0.01	20	39	0.00	31	0.02	42	0.00	29	0.05	36	0.00	26	0.18	31	0.02	29	0.05	43	0.00
	HAR (Norm)	0.01	20	33	0.01	17	0.51	32	0.01	34	0.00	10	0.01	6	0.00	20	0.97	15	0.25	27	0.13
	Lasso (Emp)	0.01	20	41	0.00	28	0.08	43	0.00	28	0.08	38	0.00	33	0.01	32	0.01	28	0.08	43	0.00
	Lasso (Norm)	0.01	20	35	0.00	15	0.25	34	0.00	35	0.00	11	0.03	7	0.00	21	0.79	15	0.25	29	0.05
	Emp	0.025	50	51	0.84	73	0.00	51	0.84	62	0.09	47	0.71	50	0.95	51	0.84	50	0.95	55	0.44
	HAR (Emp)	0.025	50	56	0.37	53	0.63	61	0.11	53	0.63	56	0.37	64	0.05	65	0.03	49	0.93	73	0.00
	HAR (Norm)	0.025	50	56	0.37	37	0.06	58	0.24	69	0.01	30	0.00	19	0.00	53	0.63	25	0.00	56	0.37
	Lasso (Emp)	0.025	50	55	0.44	49	0.93	63	0.06	57	0.30	60	0.15	67	0.02	70	0.01	51	0.84	80	0.00
	Lasso (Norm)	0.025	50	57	0.30	36	0.04	60	0.15	66	0.02	32	0.01	21	0.00	54	0.53	28	0.00	58	0.24
1000	Emp	0.01	14	35	0.00	54	0.00	26	0.00	40	0.00	23	0.02	32	0.00	27	0.00	21	0.07	21	0.07
	HAR (Emp)	0.01	14	27	0.00	27	0.00	22	0.04	21	0.07	22	0.04	24	0.01	18	0.28	7	0.04	19	0.19
	HAR (Norm)	0.01	14	19	0.19	10	0.28	22	0.04	25	0.01	4	0.00	3	0.00	15	0.76	3	0.00	19	0.19
	Lasso (Emp)	0.01	14	26	0.00	25	0.01	25	0.01	19	0.19	21	0.07	25	0.01	18	0.28	8	0.09	20	0.12
	Lasso (Norm)	0.01	14	21	0.07	10	0.28	23	0.02	22	0.04	3	0.00	2	0.00	14	0.96	2	0.00	20	0.12
	Emp	0.025	35	66	0.00	105	0.00	48	0.03	75	0.00	47	0.04	57	0.00	57	0.00	43	0.16	43	0.16
	HAR (Emp)	0.025	35	39	0.46	44	0.12	37	0.68	32	0.65	34	0.92	55	0.00	38	0.56	24	0.05	39	0.46
	HAR (Norm)	0.025	35	33	0.78	24	0.05	40	0.36	47	0.04	16	0.00	12	0.00	32	0.65	9	0.00	42	0.22
	Lasso (Emp)	0.025	35	43	0.16	43	0.16	35	0.94	38	0.56	34	0.92	50	0.01	38	0.56	24	0.05	39	0.46
	Lasso (Norm)	0.025	35	34	0.92	23	0.03	37	0.68	52	0.01	14	0.00	13	0.00	32	0.65	10	0.00	40	0.36
2000	Emp	0.01	4	0	–	0	–	1	0.08	0	–	0	–	2	0.30	0	–	1	0.08	0	–
	HAR (Emp)	0.01	4	6	0.30	0	–	6	0.30	6	0.30	5	0.57	6	0.30	5	0.57	3	0.66	4	0.93
	HAR (Norm)	0.01	4	1	0.08	0	–	5	0.57	5	0.57	0	–	1	0.08	1	0.08	0	–	2	0.30
	Lasso (Emp)	0.01	4	6	0.30	0	–	6	0.30	6	0.30	5	0.57	6	0.30	6	0.30	2	0.30	4	0.93
	Lasso (Norm)	0.01	4	1	0.08	0	–	5	0.57	5	0.57	0	–	2	0.30	0	–	0	–	2	0.30
	Emp	0.025	10	10	0.89	4	0.04	5	0.10	6	0.21	1	0.00	5	0.10	3	0.01	3	0.01	1	0.00
	HAR (Emp)	0.025	10	7	0.38	0	–	11	0.65	7	0.38	11	0.65	10	0.89	10	0.89	4	0.04	7	0.38
	HAR (Norm)	0.025	10	4	0.04	0	–	8	0.60	8	0.60	0	–	3	0.01	5	0.10	1	0.00	4	0.04
	Lasso (Emp)	0.025	10	7	0.38	0	–	11	0.65	7	0.38	11	0.65	9	0.85	10	0.89	4	0.04	8	0.60
	Lasso (Norm)	0.025	10	5	0.10	0	–	7	0.38	7	0.38	0	–	4	0.04	6	0.21	1	0.00	5	0.10

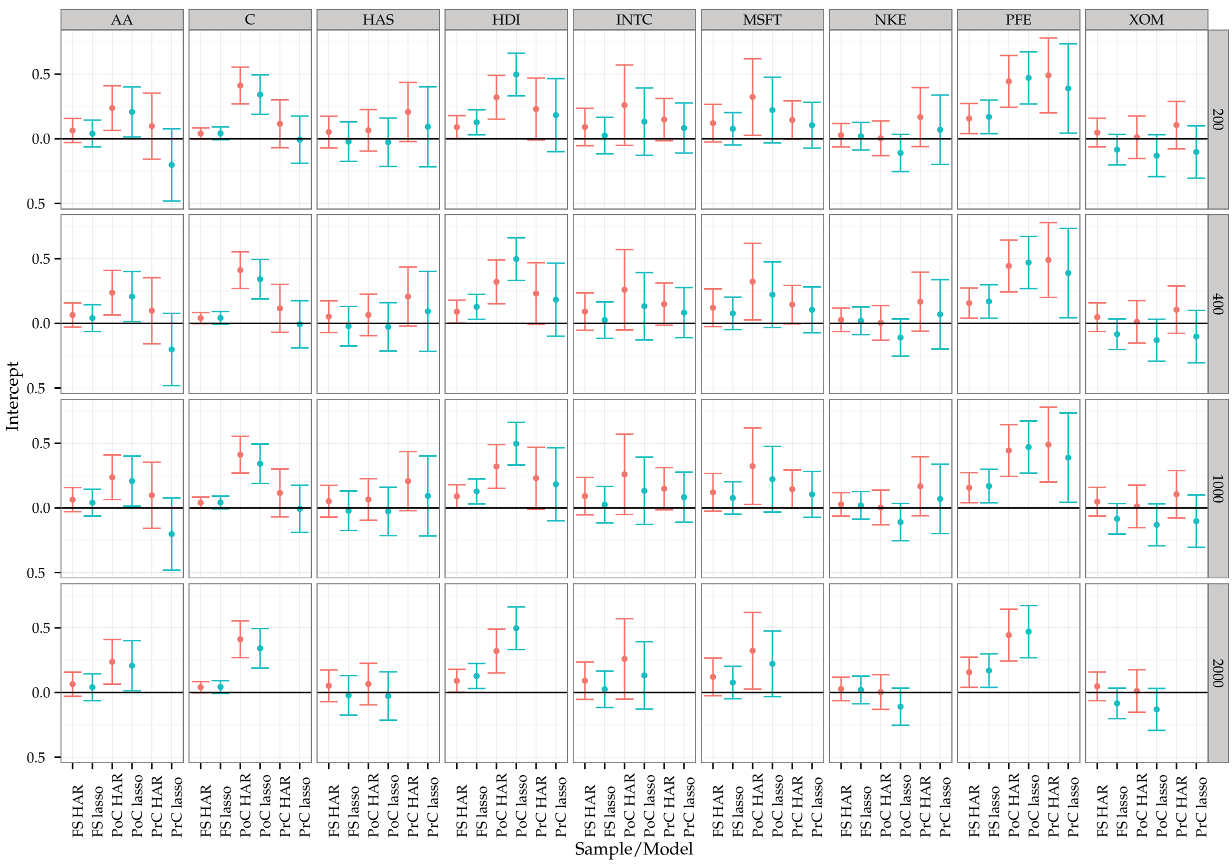
Figure A.5: Kernel density estimates of standardized log-returns for pre-crisis (PC) and full sample (FS) against normal distribution.



A.5 Mincer Zarnowitz Regressions

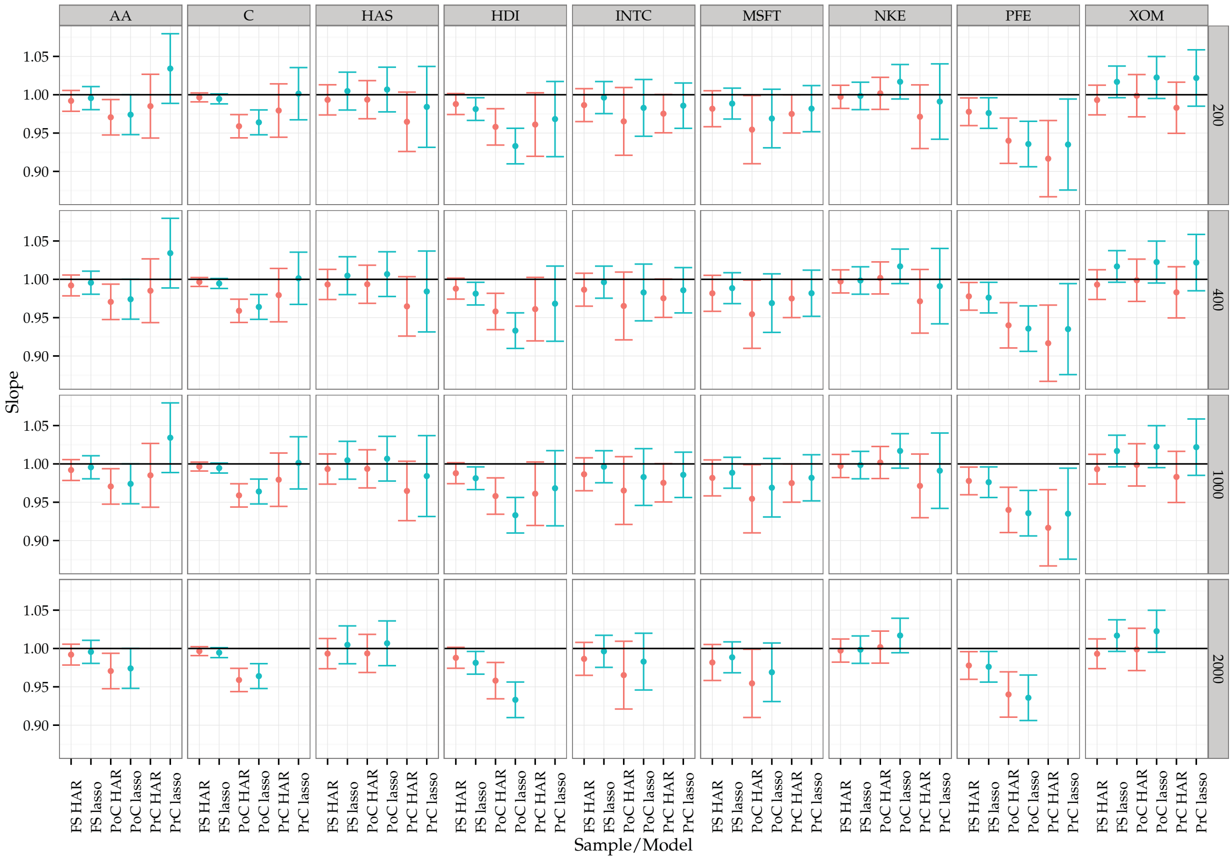
In this paragraph we present the Mincer-Zarnowitz (Mincer & Zarnowitz 1969) regressions for the lasso as well as the HAR model for the different training window lengths as well as split into pre-crisis (PrC), post-crisis (PoC), and full-sample (FS). Instead of reporting tables we include three figures: Fig. A.6 contains the estimated intercept with 95% confidence intervals, Fig. A.7 contains the estimated slope parameter with 95% confidence intervals, and Fig. A.8 contains the p-value of the joint hypothesis that the intercept equals 0 and the slope equals 1. Horizontal lines show the 5% and 10% level. In total the lasso is rejected 38 times (48 times) at the 5% level (10% level) whereas the HAR is rejected 50 times in both cases (out of 99 tests for each model). We account for dependence of the error term by using HAC consistent standard errors (Newey & West 1987).

Figure A.6: Mincer Zarnowitz Estimates (Intercept)



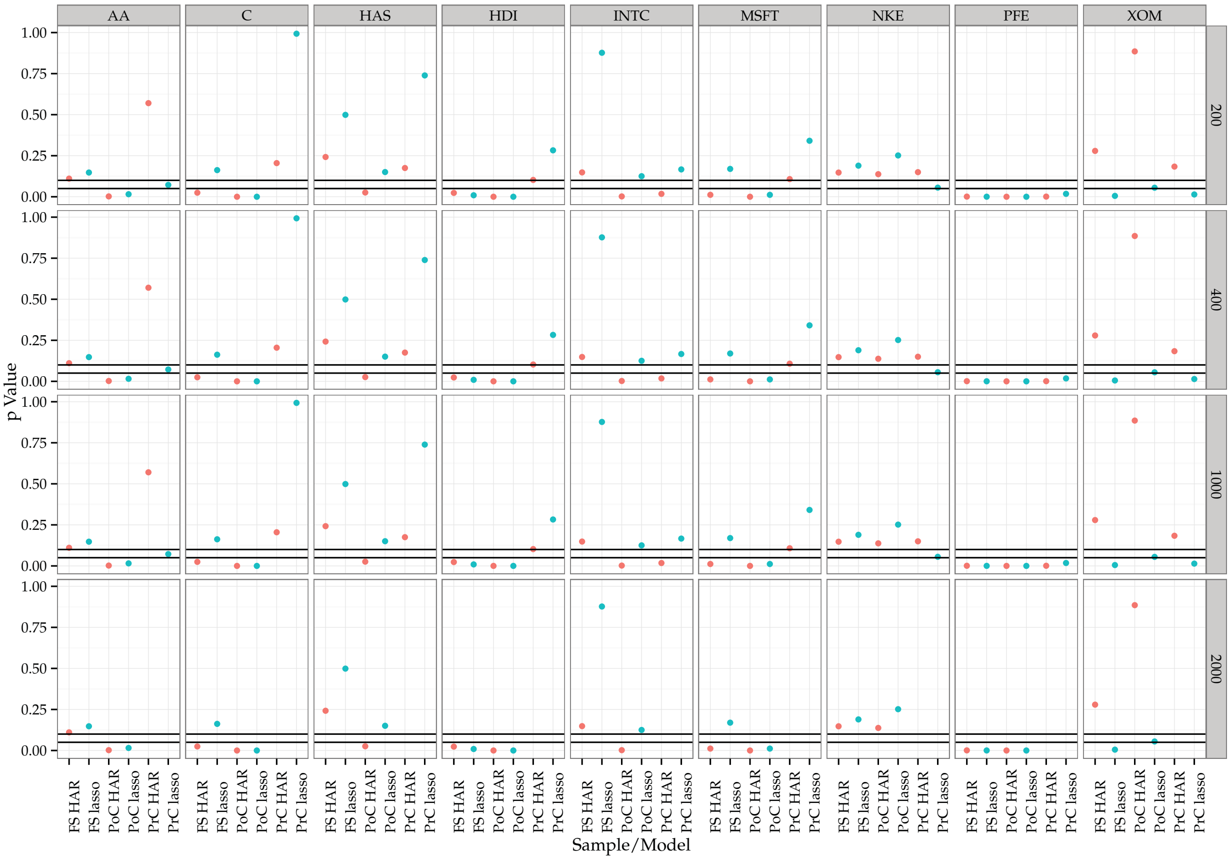
Estimate of $\hat{\alpha}$ in $\log \mathrm{RV}_t = \alpha + \beta \cdot \widehat{\log \mathrm{RV}}_t + \epsilon_t$

Figure A.7: Mincer Zarnowitz Estimates (Slope)



Estimate of $\hat{\beta}$ in $\log \mathrm{RV}_t = \alpha + \beta \cdot \widehat{\log \mathrm{RV}}_t + \epsilon_t$

Figure A.8: Joint hypothesis p-values



p-value $\mathcal{H}_0 : \alpha = 0 \wedge \beta = 1$ of $\log \text{RV}_t = \alpha + \beta \cdot \widehat{\log \text{RV}}_t + \epsilon_t$

A.6 Implementation

The present paper uses R, the statistical programming language (R Core Team 2012) in its version 2.14.1 The lasso estimates were obtained using the `glmnet` package which is based on Friedman et al. (2010) as well as the `lars` package (Hastie & Efron 2011). All codes used herein are available from the author upon request.

Appendix B

Markets from East to West, News and Volatility: Comparing Forecast Accuracy

B.1 Data Description

B.1.1 Market Data

I use five-minute price data from tickdata.com spanning Jan 2, 2007 to Sep 25, 2012 of the three cash indices S&P 500, FTSE 100, and Nikkei 225, each featuring two decimals.

The S&P 500 (Standard & Poors) is an index containing 500 large cap US-traded companies and is weighted according to market capitalization. Its constituents span different sectors (Poors 2013) denoted in US dollars.

The FTSE (Financial Times Stock Exchange) is an index comprising 100 UK-traded blue chip companies from different sectors and is denoted in British pounds (FTSE Group 2013).

The Nikkei 225 is a price-weighted index constituted of 225 stocks listed on the on the Tokyo Stock Exchange (Nikkei 2013) comprising different sectors and is denoted in Japanese yen.

The data are available on a trading day from 9.30 a.m. to 4.00 p.m. (S&P 500, local time), 8.00 a.m. to 4.30 p.m. (FTSE 1000, local time), and 9.00 a.m. to 3.00 p.m. (Nikkei 225, local time).[1]

B.1.2 Volatility Estimators

In the following I briefly review the estimators used in the current application. An overview this topic might also be found in McAleer & Medeiros (2008) or in Bauwens, Hafner & Laurent (2012). Altough *realized variance* is often said to be a model-free estimator its meaningfulness still hinges on the common assumption of a (jump-) diffusion. Let an asset's log price p_t be governed by the following process:

$$\mathrm{d}p_t = \mu_t \mathrm{d}t + \sigma_t \mathrm{d}W_t$$

where the mean process μ_t is continuous and of finite variation, σ_t is the instantaneous volatility which is assumed to be càdlàg and W_t is a standard Brownian motion.

[1]Tokyo Stock Exchange has a break from 11.30 a.m. to 12.30 p.m. Prior to Nov 21, 2011 the break was from 11.00 a.m. to 12.30 p.m. For the sake of consistency I maintain the fictitious break from 11.00 a.m. to 12.30 p.m. through the whole sample.

Then, if one defines the log *return* of a day (with a day normalized to 1), $r_t = p_t - p_{t-1}$, one finds that

$$r_t = \int_{t-1}^{t} \mu_s \mathrm{d}s + \int_{t-1}^{t} \sigma_s \mathrm{d}W_s$$

such that

$$r_t \sim \mathcal{N}\left(\int_{t-1}^{t} \mu_s \mathrm{d}s, \int_{t-1}^{t} \sigma_s^2 \mathrm{d}s\right)$$

where $\int_{t-1}^{t} \sigma_s^2 \mathrm{d}s$ is commonly denoted as *integrated volatility*, IV_t.

If this process is now observed at M discrete times $t_j \in [t-1, t]$ for a specific day t one may compute *intraday returns* $r_{t_j} = p_{t_j} - p_{t_{j-1}}$. One can then show that *realized volatility*, $\mathrm{RV}_t = \sum_{j=1}^{M} r_{t_j}^2$, converges to IV_t, i.e., if $\Delta = \max_i t_i - t_{i-1}$, then $\mathrm{RV}_t \operatorname{plim}_{\Delta \to 0} \mathrm{IV}_t$. Since high-frequency asset prices are often plagued by microstructure noise or jump components in the price process, this standard assumption may not be sufficient. However, since the data are index data, I consider microstructure noise to be of minor importance and focus on possible jumps. If one thus relaxes the assumption of the log price process following an Itō process to the class of Brownian semi-martingale with finite activity jumps (BSMFAJ) one may show that RV_t no longer converges to IV_t such that an alternative is required. One possible estimator introduced is the *bipower variation*, defined as

$$\mathrm{BPV}_t = \mu^{-2} \frac{M}{M-1} \sum_{i=2}^{M} |r_{t_i}||r_{t_{i-1}}|, \qquad \mu = \sqrt{2/\pi},$$

which can be shown to converge to IV_t under the assumption of BSMFAJ (Barndorff-Nielsen & Shephard 2004, Barndorff-Nielsen, Shephard & Winkel 2006).

By slight abuse of language I will refer to *realized volatility* as the quantitiy computed by either RV_t or BPV_t and provide further information to avoid confusion when needed.

To compute the realized volatilities of each of the three indices I use a five-minute aggregation of prices p_t such that $t_j - t_{j-1} = 300$ seconds and use the bipower variation as a primary estimator for the integrated volatility.[2] As an alternative measure I have also included the results for the naive quadratic variation estimator RV_t in Appendix B.2. As the Nikkei is traded in two separate sessions a day, I compute RV_{t_m} for the morning session, and RV_{t_a} for the afternoon session and add these two as $\mathrm{RV}_t = \mathrm{RV}_{t_m} + \mathrm{RV}_{t_a}$ as suggested in Ishida & Toshiaki (2009).

Since for the multivariate application cannot simply omit missing values I impute the values of RV_{t-1} if the observation on date $t-1$ is missing. For the out-of-sample evaluation I do not impute any values for the S&P's realized volatility to avoid distortion of the out-of-sample results.

As is common practice I then transform RV_t series by the logarithm to obtain a time series that is amenable to modeling: $\log \mathrm{RV}_t$ features near-normality in terms of unconditional distribution, further more, this transform ensures positiveness of volatility estimates as well as increasing the variance explained for a linear volatility model (Audrino & Knaus 2012). In the following I hence provide descriptive statistics for each of the three $\log \mathrm{RV}_t$ series. Although I follow a strict separation of in-sample (Jan 3, 2007 to Dec 31, 2010) and out-of-sample (Jan 1, 2011 to Sep 25, 2012) I report the descriptive statistics for the whole sample jointly in Tab. B.1 in the interest of space. Fig. B.2 contains a graphical illustration of the autocorrelations found in the sample.

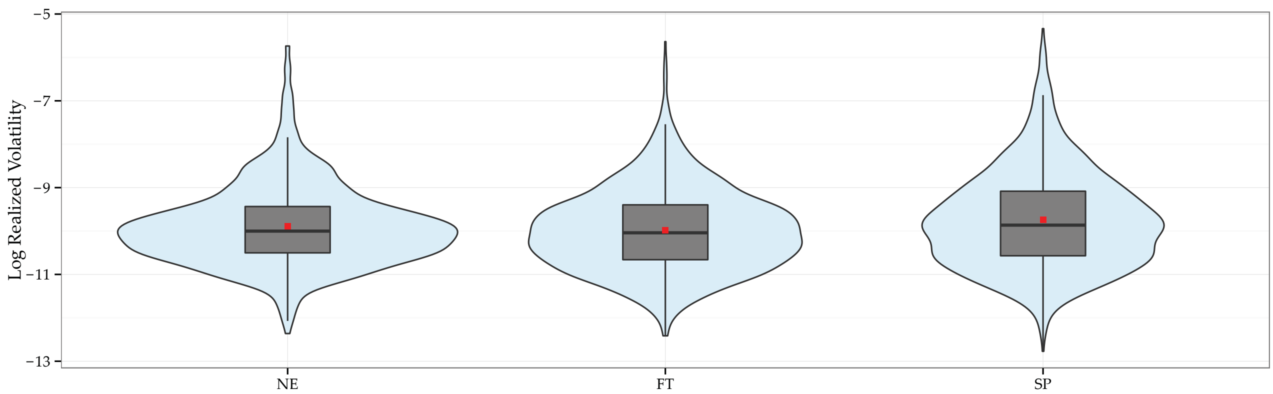
[2] To mitigate the stale quote problem the first 5 minute return is discarded, thus leaving $M = 77$ 5-minutes returns for both S&P 500 and FTSE 1000 (pseudo-closing), and $M = 53$ 5-minutes returns for the Nikkei 225.

Table B.1: Descriptive statistics

	NE	FT	SP
Observations	1406	1417	1445
Mean	-9.889	-9.979	-9.739
Standard Dev.	0.936	0.955	1.132
Ex. Kurtosis	1.880	0.629	0.476
Skewness	0.923	0.532	0.631
ρ_1	0.795	0.811	0.865
ρ_{10}	0.618	0.622	0.674
ρ_{20}	0.513	0.499	0.551
ρ_{50}	0.337	0.331	0.369

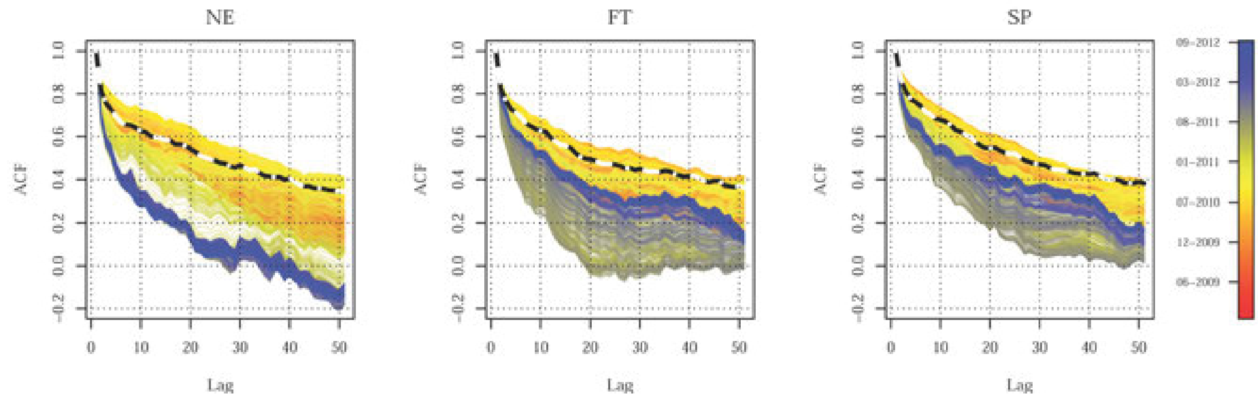
Descriptive statistics for the three $\log \widehat{\mathrm{RV}}_t$ series using bipower variation. ρ_i denotes the autocorrelation at lag i.

Figure B.1: Distribution of unconditional $\log \mathrm{RV}_t$



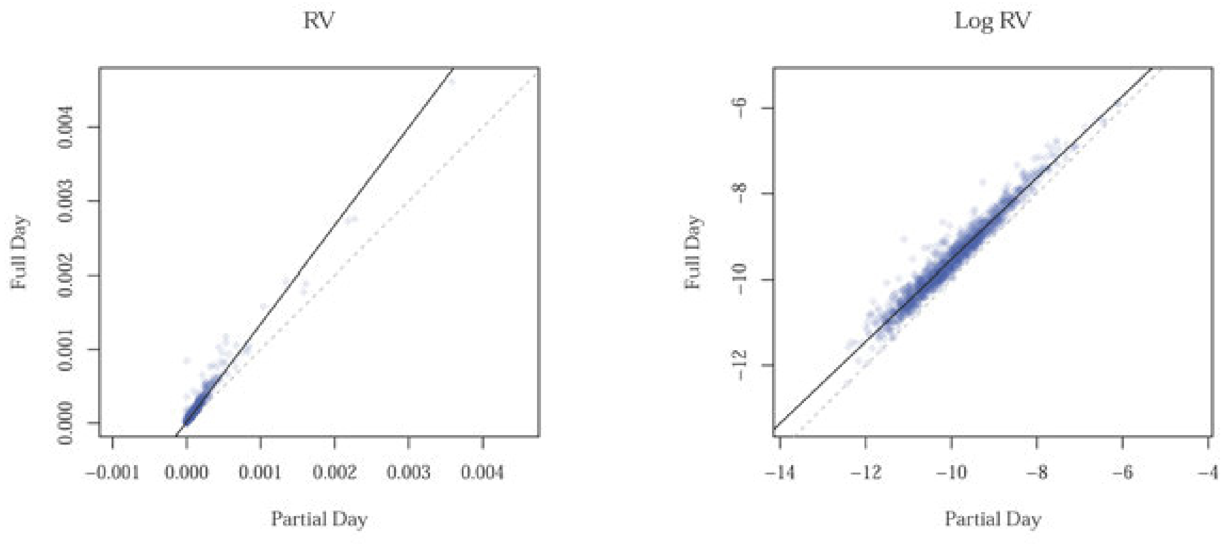
Violin plots (Hintze & Nelson 1998) of the unconditional $\log \mathrm{RV}_t$ distribution for the three markets with an inset boxplot. The mean is indicated as a red dot.

Figure B.2: Autocorrelation over time



ACF over time with a rolling width of 500 observations. The end date of window is color coded according to the scheme on the right-hand side. The total sample ACF is shown as a black dashed line.

Figure B.3: Full day versus partial day realized volatility



Regression plot of full day (log) volatility (ordinate) versus pseudo-closing (log) volatility (abscissa) for the FT with inset least square fit (solid black) and bisectrix (dashed gray).

Table B.2: Regression estimates full versus partial day

	Realized Volatility	Log Realized Volatility
Intercept	0.00002	0.01001
	(0.00000)	(0.07386)
Slope	1.32656	0.95497
	(0.02956)	(0.00743)
R^2	0.964	0.953
N	1417	1417

Regression estimates of $x_{\text{Full Day}} \sim \alpha + \beta \cdot x_{\text{Partial Day}}$ with Newey-West adjusted standard errors for realized volatility as well as log realized volatility.

B.1.3 News Data

The news data I used are collected by forexfactory.com which serves as an aggregator of different news sources. Relevant to the present study are the following sources and news events: GDP (Bureau of Economic Analysis, final GDP), unemployment figures (Bureau of Labor Statistics, unemployment rate), inflation (Bureau of Labor Statistics, CPI), and federal fund rates (Federal Reserve). The release schedule for these indicators as used in the present study is collected below.

Employment

Jan 04, 2007, Feb 02, 2007, Mar 09, 2007, Apr 06, 2007, May 04, 2007, Jun 01, 2007, Jul 06, 2007, Aug 03, 2007, Sep 07, 2007, Oct 05, 2007, Nov 02, 2007, Dec 07, 2007, Jan 04, 2008, Feb 01, 2008, Mar 07, 2008, Apr 04, 2008, May 02, 2008, Jun 06, 2008, Jul 03, 2008, Aug 01, 2008, Sep 05, 2008, Oct 03, 2008, Nov 07, 2008, Dec 05, 2008, Jan 09, 2009, Feb 06, 2009, Mar 06, 2009, Apr 03, 2009, May 08, 2009, Jun 05, 2009, Jul 02, 2009, Aug 07, 2009, Sep 04, 2009, Oct 02, 2009, Nov 06, 2009, Dec 04, 2009, Jan 08, 2010, Feb 05, 2010, Mar 05, 2010, Apr 02, 2010, May 07, 2010, Jun 04, 2010, Jul 02, 2010, Aug 06, 2010, Sep 03, 2010, Oct 08, 2010, Nov 05, 2010, Dec 03, 2010, Jan 07, 2011, Feb 04, 2011, Mar 04, 2011, Apr 01, 2011, May 06, 2011, Jun 03, 2011, Jul 08, 2011, Aug 05, 2011, Sep 02, 2011, Oct 07, 2011, Nov 04, 2011, Dec 02, 2011, Jan 06, 2012, Feb 03, 2012, Mar 09, 2012, Apr 06, 2012, May 04, 2012, Jun 01, 2012, Jul 06, 2012, Aug 03, 2012, Sep 07, 2012

GDP

Jan 31, 2007, Feb 28, 2007, Mar 29, 2007, Apr 27, 2007, May 31, 2007, Jun 28, 2007, Jul 27, 2007, Aug 30, 2007, Sep 27, 2007, Oct 31, 2007, Nov 29, 2007, Dec 20, 2007, Jan 30, 2008, Feb 28, 2008, Mar 27, 2008, Apr 30, 2008, May 29, 2008, Jun 26, 2008, Jul 31, 2008, Aug 28, 2008, Sep 26, 2008, Oct 30, 2008, Nov 25, 2008, Dec 23, 2008, Jan 30, 2009, Feb 27, 2009, Mar 26, 2009, Apr 29, 2009, May 29, 2009, Jun 25, 2009, Jul 31, 2009, Aug 27, 2009, Sep 30, 2009, Oct 29, 2009, Nov 24, 2009, Dec 22, 2009, Jan 29, 2010, Feb 26, 2010, Mar 26, 2010, Apr 30, 2010, May 27, 2010, Jun 25, 2010, Jul 30, 2010, Aug 27, 2010, Sep 30, 2010, Oct 29, 2010, Nov 23, 2010, Dec 22, 2010, Jan 28, 2011, Feb 25, 2011, Mar 25, 2011, Apr 28, 2011, May 26, 2011, Jun 24, 2011, Jul 29, 2011, Aug 26, 2011, Sep 29, 2011, Oct 27, 2011, Nov 22, 2011, Dec 22, 2011, Jan 27, 2012, Feb 29, 2012, Mar 29, 2012, Apr 27, 2012, May 31, 2012, Jun 28, 2012, Jul 27, 2012, Aug 29, 2012

CPI

Jan 18, 2007, Feb 21, 2007, Mar 16, 2007, Apr 17, 2007, May 15, 2007, Jun 15, 2007, Jul 18, 2007, Aug 15, 2007, Sep 19, 2007, Oct 17, 2007, Nov 15, 2007, Dec 14, 2007, Jan 16, 2008, Feb 20, 2008, Mar 14, 2008, Apr 16, 2008, May 14, 2008, Jun 13, 2008, Jul 16, 2008, Aug 14, 2008, Sep 16, 2008, Oct 16, 2008, Nov 19, 2008, Dec 16, 2008, Jan 16, 2009, Feb 20, 2009, Mar 18, 2009, Apr 15, 2009, May 15, 2009, Jun 17, 2009, Jul 15, 2009, Aug 14, 2009, Sep 16, 2009, Oct 15, 2009, Nov 18, 2009, Dec 16, 2009, Jan 15, 2010, Feb 19, 2010, Mar 18, 2010, Apr 14, 2010, May 19, 2010, Jun 17, 2010, Jul 16, 2010, Aug 13, 2010, Sep 17, 2010, Oct 15, 2010, Nov 17, 2010, Dec 15, 2010, Jan 14, 2011, Feb 17, 2011, Mar 17, 2011, Apr 15, 2011, May 13, 2011, Jun 15, 2011, Jul 15, 2011, Aug 18, 2011, Sep 15, 2011, Oct 19, 2011, Nov 16, 2011, Dec 16, 2011, Jan 19, 2012, Feb 17, 2012, Mar 16, 2012, Apr 13, 2012, May 15, 2012, Jun 14, 2012, Jul 17, 2012, Aug 15, 2012, Sep 14, 2012

FED

Jan 31, 2007, Mar 21, 2007, May 09, 2007, Jun 28, 2007, Aug 07, 2007, Sep 18, 2007, Oct 31, 2007, Dec 11, 2007, Jan 22, 2008, Jan 30, 2008, Mar 18, 2008, Apr 30, 2008, Jun 25, 2008, Aug 05, 2008, Sep 16, 2008, Oct 08, 2008, Oct 29, 2008, Dec 16, 2008, Jan 28, 2009, Mar 18, 2009, Apr 29, 2009, Jun 24, 2009, Aug 12, 2009, Sep 23, 2009, Nov 04, 2009, Dec 16, 2009, Jan 27, 2010, Mar 16, 2010, Apr 28, 2010, Jun 23, 2010, Aug 10, 2010, Sep 21, 2010, Nov 03, 2010, Dec 14, 2010, Jan 26, 2011, Mar 15, 2011, Apr 27, 2011, Jun 22, 2011, Aug 09, 2011, Sep 21, 2011, Nov 02, 2011, Dec 13, 2011, Jan 25, 2012, Mar 13, 2012, Apr 25, 2012, Jun 20, 2012, Aug 01, 2012, Sep 13, 2012

B.2 Robustness

The subsequent paragraphs contain a sensitivity analysis with regard to training window length as well as a robustness analysis when using the naive quadratic estimator to compute realized volatility instead of the bipower variation estimator.

B.2.1 Sample Length Sensitivity

Table B.3: Out-of-sample forecasts with varying training window widths

					100			200			500			1000		
	δ_1	δ_2	δ_3	δ_4	MSPE	MAPE	R^2	MSPE	MAPE	R^2	MSPE	MAPE	R^2	MSPE	MAPE	R^2
	1	0	0	-	0.315**	0.437***	0.642	0.287*	0.422*	0.661	0.279	0.421	0.668	0.281	0.422	0.670
	0	1	0	-	0.317	0.435**	0.637	0.292	0.424	0.654	0.285	0.424	0.662	0.285	0.423	0.665
	1	1	0	-	0.315**	0.434***	0.643	0.282**	0.414**	0.666	0.273	0.416	0.675	0.274	0.416	0.678
C_1	0	0	1	-	0.307**	0.427***	0.647	0.284*	0.418**	0.664	0.274	0.418	0.674	0.278	0.419	0.673
	1	0	1	-	0.307**	0.435***	0.650	0.275**	0.417**	0.674	0.265*	0.413	0.684	0.270	0.416	0.682
	0	1	1	-	0.307**	0.426***	0.647	0.280*	0.415**	0.668	0.271	0.415	0.677	0.275	0.416	0.676
	1	1	1	-	0.307**	0.431***	0.651	0.272***	0.410**	0.678	0.261**	0.407*	0.689	0.264	0.410	0.688
	0	0	0	0	0.318	0.439*	0.634	0.301	0.431	0.643	0.291	0.427	0.655	0.290	0.426	0.658
	1	0	0	0	0.315*	0.441*	0.638	0.293*	0.426	0.653	0.281	0.424	0.667	0.281	0.423	0.669
	0	1	0	0	0.293**	0.427***	0.663	0.273**	0.417*	0.677	0.260*	0.410	0.692	0.264	0.411	0.689
	1	1	0	0	0.258***	0.404***	0.708	0.240***	0.389***	0.719	0.228***	0.379***	0.735	0.231***	0.381***	0.732
	0	0	1	0	0.297**	0.426***	0.657	0.281*	0.419*	0.667	0.269	0.414	0.681	0.273	0.416	0.678
	1	0	1	0	0.263***	0.406***	0.701	0.246***	0.391***	0.712	0.237***	0.383***	0.725	0.239**	0.386**	0.722
	0	1	1	0	0.313	0.437**	0.640	0.292	0.426	0.654	0.281	0.421	0.666	0.281	0.419	0.670
	1	1	1	0	0.289**	0.422***	0.668	0.265***	0.406**	0.687	0.253**	0.402	0.700	0.255	0.403	0.699
C_2	0	0	0	1	0.266***	0.408***	0.697	0.256***	0.397**	0.701	0.247**	0.391*	0.714	0.247	0.392	0.714
	1	0	0	1	0.261***	0.406***	0.703	0.250***	0.396**	0.708	0.238**	0.388*	0.724	0.239**	0.388*	0.724
	0	1	0	1	0.260***	0.403***	0.705	0.244***	0.389***	0.715	0.232***	0.380***	0.731	0.232***	0.379***	0.732
	1	1	0	1	0.294**	0.422***	0.661	0.272**	0.411**	0.677	0.262	0.407	0.689	0.265	0.409	0.688
	0	0	1	1	0.261***	0.403***	0.703	0.248***	0.388***	0.711	0.238***	0.382***	0.725	0.239**	0.383***	0.724
	1	0	1	1	0.311**	0.436***	0.645	0.283**	0.418**	0.665	0.272	0.417	0.678	0.271	0.414	0.682
	0	1	1	1	0.262***	0.403***	0.705	0.240***	0.384***	0.720	0.230***	0.376***	0.733	0.232***	0.378***	0.731
	1	1	1	1	0.259***	0.403***	0.709	0.236***	0.384***	0.724	0.224***	0.373***	0.740	0.225***	0.373***	0.739
C_3	-	-	-	-	0.264***	0.409***	0.693	0.248***	0.392***	0.708	0.233***	0.382***	0.730	0.232***	0.380***	0.732
C_4	-	-	-	-	0.273***	0.415***	0.678	0.268**	0.413**	0.684	0.254**	0.403	0.701	0.248	0.398	0.710

Out-of-sample forecasts using bipower estimator for RV_t with varying training window widths. The table is equivalent to Tab. 2.3 with the exception that stars for R^2 are omitted for typographical ease.

B.2.2 Out-of-sample Results Naive Estimator

Table B.4: Out-of-sample results using the naive estimator

	δ_1	δ_2	δ_3	δ_4	MSPE	Var	Bias	MAPE	R^2
C_1	1	0	0	-	0.263	0.263	−0.011%	0.409	0.677
	0	1	0	-	0.270	0.270	−0.013%	0.414	0.670
	1	1	0	-	0.259	0.259	−0.009%	0.407	0.683
	0	0	1	-	0.260	0.260	−0.011%	0.406	0.680
	1	0	1	-	0.252	0.253	−0.008%	0.402	0.690
	0	1	1	-	0.259	0.259	−0.009%	0.406	0.682
	1	1	1	-	0.249	0.249	−0.006%	0.399	0.694
C_2	0	0	0	0	0.272	0.273	−0.039%	0.414	0.666
	1	0	0	0	0.264	0.264	−0.012%	0.411	0.677
	0	1	0	0	0.247	0.247	0.038%	0.398	0.696
	1	1	0	0	0.218***	0.216	0.315%	0.370***	0.736**
	0	0	1	0	0.256	0.255	0.015%	0.402	0.686
	1	0	1	0	0.226*	0.225	0.299%	0.376***	0.725**
	0	1	1	0	0.266	0.266	−0.031%	0.410	0.675
	1	1	1	0	0.241	0.241	0.038%	0.392	0.704
	0	0	0	1	0.234	0.233	0.298%	0.382	0.717**
	1	0	0	1	0.225*	0.223	0.315%	0.375***	0.728**
	0	1	0	1	0.220***	0.218	0.308%	0.368***	0.734**
	1	1	0	1	0.250	0.250	0.019%	0.397	0.693
	0	0	1	1	0.228*	0.227	0.298%	0.376***	0.724**
	1	0	1	1	0.255	0.256	−0.009%	0.404	0.687
	0	1	1	1	0.221***	0.219	0.300%	0.370***	0.732**
	1	1	1	1	0.213***	0.212	0.310%	0.364***	0.741**
C_3	-	-	-	-	0.220***	0.218	0.380%	0.370***	0.734***
C_4	-	-	-	-	0.240	0.238	0.318%	0.387	0.710**

Out-of-sample forecast results with naive RV_t. Otherwise equivalent to Tab. 2.3

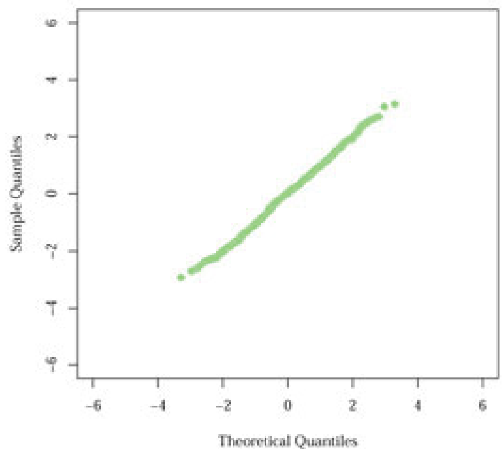
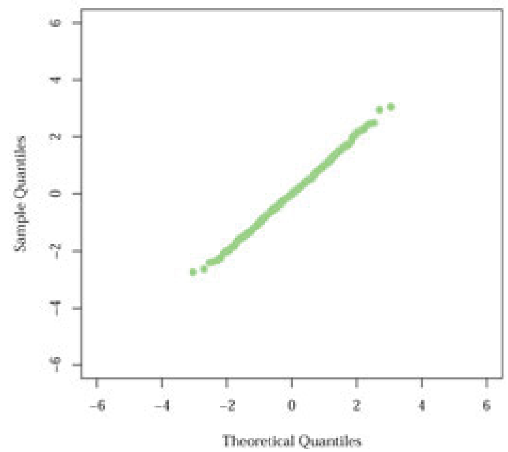
B.2.3 Value-at-risk Results Naive Estimator

Table B.5: Out-of-sample forecasts with varying training window widths using naive estimator

					100			200			500			1000		
	δ_1	δ_2	δ_3	δ_4	MSPE	MAPE	R^2	MSPE	MAPE	R^2	MSPE	MAPE	R^2	MSPE	MAPE	R^2
	1	0	0	-	0.294**	0.427***	0.652	0.269**	0.414	0.669	0.261	0.409	0.677	0.263	0.409	0.677
	0	1	0	-	0.297**	0.427***	0.646	0.275	0.415	0.661	0.269	0.414	0.668	0.270	0.414	0.670
	1	1	0	-	0.295**	0.427***	0.652	0.265**	0.408**	0.674	0.257	0.407	0.682	0.259	0.407	0.683
C_1	0	0	1	-	0.286**	0.417***	0.657	0.265**	0.406**	0.673	0.257	0.404	0.682	0.260	0.406	0.680
	1	0	1	-	0.285***	0.425***	0.661	0.257***	0.406**	0.684	0.248**	0.401	0.694	0.252	0.402	0.690
	0	1	1	-	0.286**	0.418***	0.657	0.262**	0.404**	0.676	0.255	0.403	0.684	0.259	0.406	0.682
	1	1	1	-	0.285***	0.424***	0.662	0.253***	0.401**	0.688	0.246**	0.398*	0.696	0.249	0.399	0.694
	0	0	0	0	0.298	0.426**	0.643	0.283	0.421	0.650	0.272	0.415	0.663	0.272	0.414	0.666
	1	0	0	0	0.298**	0.432*	0.645	0.274*	0.416	0.662	0.262	0.412	0.676	0.264	0.411	0.677
	0	1	0	0	0.276***	0.421***	0.671	0.255**	0.404**	0.685	0.243**	0.396*	0.700	0.247	0.398	0.696
	1	1	0	0	0.246***	0.394***	0.710	0.228***	0.376***	0.722	0.215***	0.368***	0.739	0.218***	0.370***	0.736
	0	0	1	0	0.279**	0.414***	0.666	0.264*	0.406**	0.674	0.252	0.399	0.689	0.256	0.402	0.686
	1	0	1	0	0.252***	0.395***	0.703	0.234***	0.379***	0.714	0.224**	0.373***	0.728	0.226*	0.376***	0.725
	0	1	1	0	0.295*	0.425***	0.648	0.276	0.417	0.659	0.265	0.410	0.672	0.266	0.410	0.675
	1	1	1	0	0.273***	0.417***	0.674	0.249***	0.397**	0.693	0.238**	0.391**	0.706	0.241	0.392	0.704
C_2	0	0	0	1	0.255***	0.396***	0.698	0.245***	0.389**	0.701	0.234**	0.382**	0.716	0.234	0.382	0.717
	1	0	0	1	0.251***	0.397***	0.704	0.238***	0.383***	0.710	0.225**	0.376***	0.728	0.225*	0.375***	0.728
	0	1	0	1	0.252***	0.395***	0.704	0.234***	0.379***	0.715	0.220***	0.370***	0.733	0.220***	0.368***	0.734
	1	1	0	1	0.276**	0.412***	0.669	0.257**	0.401**	0.683	0.246	0.395	0.695	0.250	0.397	0.693
	0	0	1	1	0.252***	0.395***	0.702	0.239***	0.384***	0.709	0.227**	0.374***	0.725	0.228*	0.376***	0.724
	1	0	1	1	0.294**	0.428***	0.651	0.266**	0.411**	0.672	0.255	0.406	0.684	0.255	0.404	0.687
	0	1	1	1	0.252***	0.397***	0.705	0.230***	0.377***	0.720	0.219***	0.368***	0.735	0.221***	0.370***	0.732
	1	1	1	1	0.249***	0.396***	0.708	0.226***	0.373***	0.725	0.212***	0.365***	0.742	0.213***	0.364***	0.741
C_3	-	-	-	-	0.255***	0.401***	0.691	0.237***	0.382***	0.709	0.221***	0.373***	0.732	0.220***	0.370***	0.734
C_4	-	-	-	-	0.269***	0.412***	0.671	0.256**	0.398**	0.685	0.242**	0.387**	0.704	0.240	0.387	0.710

Out-of-sample forecasts using naive estimator for RV_t with varying training window widths. The table is equivalent to Tab. 2.3 with the exception that stars for R^2 are omitted for typographical ease.

Figure B.4: Distribution of standardized returns

	In-sample	Out-of-sample
$\frac{r_t}{\sqrt{RV_t}}$		
Shapiro-Wilk	0.3724	0.9416
Jarque-Bera	0.8172	0.8064

Quantile-quantile plot of standardized scaled logarithmic returns against normal distribution with the p values of normality tests.

Table B.6: Out-of-sample value-at-risk forecasts using naive estimator

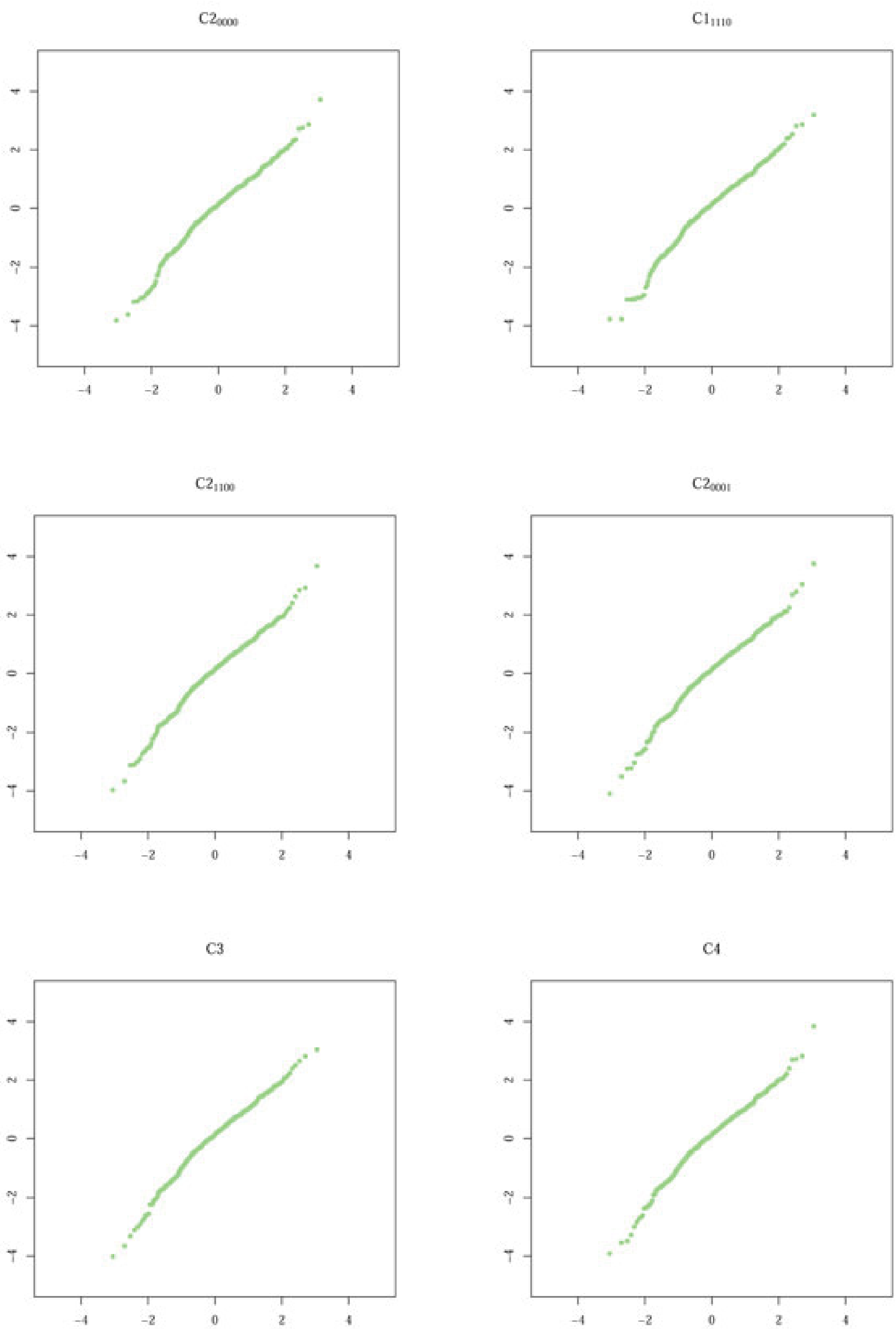
					VaR 1%				VaR 2.5%			
	δ_1	δ_2	δ_3	δ_4	Hit Ratio	UC	IND	CC	Hit Ratio	UC	IND	CC
C_1	1	0	0	-	3.20%	0.00	**0.34**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	0	1	0	-	2.97%	0.00	**0.37**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	1	1	0	-	2.97%	0.00	**0.37**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	0	0	1	-	3.20%	0.00	**0.34**	0.00	3.89%	**0.08**	**0.24**	**0.11**
	1	0	1	-	3.20%	0.00	**0.34**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	0	1	1	-	3.20%	0.00	**0.34**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	1	1	1	-	2.97%	0.00	**0.37**	0.00	4.35%	*0.02*	**0.19**	*0.03*
C_2	0	0	0	0	3.20%	0.00	**0.34**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	1	0	0	0	3.20%	0.00	**0.34**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	0	1	0	0	3.20%	0.00	**0.34**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	1	1	0	0	2.97%	0.00	**0.37**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	0	0	1	0	3.66%	0.00	**0.27**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	1	0	1	0	3.20%	0.00	**0.34**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	0	1	1	0	3.20%	0.00	**0.34**	0.00	4.12%	*0.05*	**0.21**	**0.06**
	1	1	1	0	3.20%	0.00	**0.34**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	0	0	0	1	2.52%	0.01	**0.45**	*0.02*	4.12%	*0.05*	**0.21**	**0.06**
	1	0	0	1	2.75%	0.00	**0.41**	0.01	3.89%	**0.08**	**0.24**	**0.11**
	0	1	0	1	2.75%	0.00	**0.41**	0.01	4.35%	*0.02*	**0.19**	*0.03*
	1	1	0	1	3.43%	0.00	**0.30**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	0	0	1	1	2.75%	0.00	**0.41**	0.01	4.35%	*0.02*	**0.19**	*0.03*
	1	0	1	1	2.97%	0.00	**0.37**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	0	1	1	1	2.97%	0.00	**0.37**	0.00	4.35%	*0.02*	**0.19**	*0.03*
	1	1	1	1	2.75%	0.00	**0.41**	0.01	4.12%	*0.05*	**0.21**	**0.06**
C_3	-	-	-	-	2.52%	0.01	**0.45**	*0.02*	4.12%	*0.05*	**0.21**	**0.06**
C_4	-	-	-	-	2.75%	0.00	**0.41**	0.01	3.89%	**0.08**	**0.24**	**0.11**
Static	-	-	-	-	0.23%	**0.05**	**0.92**	**0.15**	0.69%	0.00	**0.81**	*0.02*

The hit ratio (HR) together with p-values of tests of correct unconditional coverage probability (UC), independence (IND), and conditional coverage probability (CC) are shown. Non-rejection at the 0.1 (0.05) level is highlighted in boldface (italics).

B.3 Implementation

The present paper uses R, the statistical programming language (R Core Team 2012) in its version 2.14.2 The lasso estimates were obtained using `glmnet` which is based on Friedman et al. (2010). All the codes used herein are available from the author upon request.

Figure B.5: Distribution of standardized (forecasted) returns



Quantile-quantile plots of returns standardized by forecasted volatility

Appendix C

Learning from Micro-level Expert Forecasts: Real-time Data, Regression Trees, and Bagging

C.1 Additional Results

Table C.1: Bagged predictors (full sample)

	BOND		CPI		HOUSING		NGDP		TBILL		TBOND		UNEMP	
	Tree	Slct	Tree	Slct	Tree	Slct	Tree	Slct	Tree	Slct	Tree	Slct	Tree	Slct
0	0.69^{**}	0.65	0.64^{**}	0.93	0.72^{*}	0.70^{***}	0.88^{**}	0.86^{*}	1.74	0.83^{**}	0.85	0.78^{**}	0.94	0.86^{**}
1	1.04	0.83^{***}	0.96	0.98	0.80^{**}	0.92	0.93	0.83^{*}	0.56^{***}	0.70^{*}	1.01	1.00	0.71^{*}	0.88^{**}
2	0.92	1.01	0.91^{**}	0.98	1.00	1.06	0.92	0.97	0.96	0.97	0.79^{*}	0.87^{*}	0.73	1.07
3	1.25	0.85^{*}	0.64	1.01	0.90^{*}	0.45	1.03	1.02	0.89	0.85^{**}	1.29	1.07	0.81	0.91^{*}
4	1.23	0.94	1.00	1.01	1.03	0.56	0.99	0.96^{*}	0.70^{*}	0.81^{**}	1.13	0.95	1.10	1.05

MSPEs of the bagged predictors, relative to their base variants without bagging (full evaluation sample, estimation via rolling windows of size 40). One-, two- and three stars indicate rejections of equal predictive ability, against the alternative that bagged methods work better, at the ten-, five- and one percent levels (one-sided Diebold-Mariano tests).

Table C.2: Bagged predictors (pre-crisis)

	BOND		CPI		HOUSING		NGDP		TBILL		TBOND		UNEMP	
	Tree	Slct	Tree	Slct	Tree	Slct	Tree	Slct	Tree	Slct	Tree	Slct	Tree	Slct
0	0.73^{*}	0.92	0.56^{***}	0.80	0.76^{*}	0.85^{**}	0.77^{***}	0.79^{***}	1.26	0.74^{***}	0.60^{***}	0.74	0.71^{*}	0.74^{***}
1	1.14	0.81^{**}	1.07	0.88^{**}	0.96	0.89	1.20	0.68^{*}	0.43^{**}	0.52^{*}	1.10	1.03	0.59^{**}	0.70^{**}
2	0.85	1.01	0.91	0.95	0.95	0.96	1.17	0.82	0.76	1.04	0.62^{**}	0.86^{**}	0.91	0.90
3	1.41	0.91	0.92	0.96	0.86	1.27	1.01	0.98	0.74	0.85	1.39	1.06	1.34	0.82^{***}
4	1.24	1.00	1.09	0.98	0.95	1.09	1.06	0.86^{**}	0.82	0.66	1.15	1.08	1.08	0.84

MSPEs of the bagged predictors, relative to their base variants without bagging (pre-crisis evaluation sample, estimation via rolling windows of size 40). One-, two- and three stars indicate rejections of equal predictive ability, against the alternative that bagged methods work better, at the ten-, five- and one percent levels (one-sided Diebold-Mariano tests).

C.2 Detailed Methods

C.2.1 Regression Trees

In line with the notation established in Section 3.3.2 (particularly, Tab. 3.2) we derive the construction of regression trees as loosely discussed in Section 3.2.1.

Suppose we are at time T and interested in making an h-step ahead forecast. Let the forecasters who have ever submitted a forecast with horizon h in the training period of length p be denoted by TS, i.e.,

$$TS = \bigcup_{t=T-(h+p+1)}^{T-(h+1)} \mathcal{E}_{t,h}.$$

Set $k = |TS|$, the number of all forecasters in the training set. Denote by $\hat{Y}_{t,h} = (\hat{Y}^{i_1}_{t,h}, \ldots, \hat{Y}^{i_k}_{t,h})'$ with $i_1, \ldots, i_k \in TS$ the k-dimensional vector containing the individual forecasts of all forecasters in the training set.

The goal is to obtain a partition $P = \{P_1, \ldots, P_m\}$ as described in Section 3.2.1. A node, or equivalently, a cell P_k is now split into left node $P_{k,l}$ and right node $P_{k,r}$ such that the MSE decreases maximally. Let $\mathfrak{S}$ be the set of all possible splits, i.e, $\mathfrak{S} = \{1, 2, \ldots, k\} \times \mathbb{R}$ and

$P^{(n)} = P_1, \ldots, P_n$ the current partition[1] such that the optimal split $\mathfrak{s}^*$ for each cell P_k can be determined as follows:

$$\mathfrak{s}^* = \underset{\mathfrak{s}\in\mathfrak{S}}{\operatorname{argmax}}\, R(P_k) - R(P_{k,l}(\mathfrak{s})) - R(P_{k,r}(\mathfrak{s})). \tag{C.1}$$

For a given split $\mathfrak{s}$ we have $P_{k,l}(\mathfrak{s}) = P_k \cap \{x : x \in \mathbb{R}^p, x^j < \xi\}$ (where the split $\mathfrak{s}$ is obviously identified with $x^j < \xi$) and $P_{k,r}(\mathfrak{s}) = P_k \setminus P_{k,l}(\mathfrak{s})$. Denote by $\bar{Y}_{P_k}$ the average within one cell, i.e.

$$\bar{Y}_{P_k} = \frac{1}{|\{\hat{Y}_{l,h} : \hat{Y}_{l,h} \in P_k\}|} \sum_{\substack{l\in \\ \{l:\hat{Y}_{l,h}\in P_k\}}} Y^T_{l+h}.$$

such that $R(P_k)$ can be written as

$$R(P_k) = \sum_{\substack{l\in \\ \{l:\hat{Y}_{l,h}\in P_k\}}} \left(Y^T_{l+h} - \bar{Y}_{P_k}\right)^2. \tag{C.2}$$

Obviously (C.2) applies to $P_{k,l}(\mathfrak{s})$ and $P_{k,r}(\mathfrak{s})$ in the same way such that (C.1) is well defined. Replace P_k in the partition $P^{(n)}$ with $P_{k,l}$ and $P_{k,r}$ for each k to obtain a new partition $P^{(2n)}$ and start over again.

Once a maximal tree (or equivalently, a maximal partition) is obtained one starts deleting leaves by cost-complexity pruning: define $R(\widehat{tr})$ to be the tree's MSE, i.e. if $\widehat{tr}$ has the partition $P^{(n)} = \{P_1, \ldots, P_n\}$ define $R(\widehat{tr})$ as follows

$$R(\widehat{tr}) = \frac{1}{p}\sum_{i=1}^{n} R(P_i).$$

Next, define $R_\alpha(\widehat{tr}) = R(\widehat{tr}) + \alpha \cdot \text{size}(\widehat{tr}) = R(\widehat{tr}) + \alpha \cdot n$. It can the be shown (Ripley 2008) that for each α there is a unique smallest subtree that minimizes R_α. α acts as a tuning parameter between goodness of fit and size of the tree. The optimal α is then determined

[1] To start set $n = 1$ such that $P = \{P_1\} = \{\mathbb{R}\}$

via cross-validation (10-fold in our application) and the final tree is obtained with the estimated optimal $\hat{\alpha}$ giving a tree with partition $P = \{P_1, \ldots, P_m\}$. Full details can be found in Breiman et al. (1984) or Ripley (2008).

C.2.2 Previous Best

Let again TS be the set of all forecasters who have submitted a forecast in the training period. Define for a forecaster i the individual MSE as

$$\text{MSE}_i = \frac{1}{p} \sum_{l=1}^{p} \left(\hat{Y}^i_{T-(l+h),h} - Y^T_{T-l} \right)^2 \text{ for } i \in TS$$

and choose the forecaster as the one with lowest MSE_i, i.e.,

$$i^* = \arg\min_{i \in TS} \text{MSE}_i$$

Forecaster i^* is then selected for the forecast in question.

C.3 Implementation

The procedures described herein were all implemented in R. The statistical open-source program is freely available at `http://cran.r-project.org/`. For the implementation of regression trees we used the package *rpart* (Therneau, Atkinson & Ripley 2010) with R 2.10.1 (x86_64-pc-linux-gnu). All codes and alternative robustness checks are available upon request.

Table C.3: Pre-crisis evaluation sample

		Forecast Combination Methods (in SPA set)						
		Tree	*TreeBagg*	*Slct*	*SlctBagg*	*CapTim*	*EW*	*AR*
BOND	0	1.14 *(0.04)*	0.83 *(0.20)*	0.68 *(0.42)*	0.62 *(0.96)*	0.84 *(0.09)*	1.00 *(0.05)*	2.03
	1	0.82 *(0.96)*	0.94 *(0.12)*	1.34 *(0.00)*	1.08 *(0.09)*	0.97 *(0.00)*	— *(0.23)*	0.78
	2	0.89 *(0.44)*	0.75 *(0.94)*	0.98 *(0.37)*	0.99 *(0.44)*	0.82 *(0.59)*	— *(0.34)*	0.72
	3	0.77 *(0.91)*	1.08 *(0.01)*	1.12 *(0.14)*	1.02 *(0.20)*	1.05 *(0.05)*	— *(0.10)*	0.78
	4	0.83 *(0.94)*	1.03 *(0.02)*	1.00 *(0.16)*	1.00 *(0.23)*	0.90 *(0.32)*	— *(0.20)*	0.82
CPI	0	1.76 *(0.00)*	0.99 *(0.14)*	1.00 *(0.24)*	0.81 *(0.89)*	0.99 *(0.08)*	1.00 *(0.04)*	1.86
	1	0.96 *(0.89)*	1.03 *(0.22)*	1.17 *(0.01)*	1.03 *(0.25)*	0.96 *(0.88)*	— *(0.43)*	1.04
	2	1.05 *(0.24)*	0.96 *(0.71)*	0.99 *(0.26)*	0.94 *(0.95)*	1.00 *(0.65)*	— *(0.47)*	1.03
	3	0.90 *(0.14)*	0.83 *(0.91)*	1.10 *(0.00)*	1.06 *(0.01)*	0.94 *(0.02)*	— *(0.03)*	0.93
	4	0.92 *(0.10)*	1.00 *(0.20)*	1.05 *(0.32)*	1.03 *(0.40)*	0.90 *(0.96)*	— *(0.00)*	0.93
HOUSING	0	0.72 *(0.17)*	0.54 *(0.99)*	1.03 *(0.01)*	0.87 *(0.02)*	0.84 *(0.08)*	1.00 *(0.01)*	1.01
	1	0.94 *(0.49)*	0.90 *(0.98)*	1.17 *(0.14)*	1.04 *(0.68)*	1.09 *(0.00)*	— *(0.60)*	0.99
	2	1.02 *(0.58)*	0.98 *(0.72)*	1.00 *(0.41)*	0.96 *(0.99)*	1.04 *(0.30)*	— *(0.57)*	1.01
	3	1.19 *(0.17)*	1.02 *(0.40)*	0.85 *(0.95)*	1.07 *(0.09)*	1.16 *(0.30)*	— *(0.25)*	1.14
	4	1.29 *(0.02)*	1.22 *(0.04)*	0.89 *(0.86)*	0.97 *(0.07)*	1.20 *(0.05)*	— *(0.28)*	1.24
NGDP	0	1.82 *(0.00)*	1.40 *(0.02)*	1.46 *(0.00)*	1.16 *(0.14)*	1.16 *(0.20)*	1.00 *(0.99)*	1.87
	1	1.37 *(0.00)*	1.64 *(0.01)*	1.83 *(0.06)*	1.25 *(0.27)*	1.39 *(0.04)*	— *(0.85)*	1.35
	2	1.28 *(0.00)*	1.49 *(0.00)*	1.40 *(0.10)*	1.15 *(0.13)*	1.42 *(0.00)*	— *(0.91)*	1.26
	3	1.74 *(0.03)*	1.76 *(0.01)*	0.80 *(0.31)*	0.79 *(0.70)*	1.36 *(0.00)*	— *(0.00)*	1.14
	4	1.15 *(0.12)*	1.22 *(0.17)*	1.19 *(0.07)*	1.02 *(0.55)*	1.32 *(0.03)*	— *(0.66)*	1.16
TBILL	0	0.92 *(0.00)*	1.15 *(0.11)*	1.30 *(0.00)*	0.97 *(0.00)*	0.56 *(0.61)*	1.00 *(0.06)*	6.11
	1	1.70 *(0.00)*	0.72 *(0.07)*	2.45 *(0.03)*	1.28 *(0.01)*	0.60 *(0.81)*	— *(0.04)*	1.60
	2	1.59 *(0.17)*	1.21 *(0.52)*	1.08 *(0.60)*	1.13 *(0.00)*	1.09 *(0.63)*	— *(0.86)*	1.31
	3	1.97 *(0.10)*	1.45 *(0.38)*	1.45 *(0.03)*	1.23 *(0.02)*	1.29 *(0.49)*	— *(0.82)*	1.67
	4	1.28 *(0.21)*	1.06 *(0.72)*	1.54 *(0.07)*	1.01 *(0.76)*	1.71 *(0.00)*	— *(0.70)*	1.27
TBOND	0	1.13 *(0.00)*	0.68 *(0.88)*	2.26 *(0.05)*	1.68 *(0.05)*	0.90 *(0.43)*	1.00 *(0.03)*	2.80
	1	0.84 *(0.85)*	0.93 *(0.31)*	0.92 *(0.80)*	0.95 *(0.51)*	1.32 *(0.19)*	— *(0.28)*	0.80
	2	1.38 *(0.03)*	0.86 *(0.51)*	1.45 *(0.02)*	1.24 *(0.14)*	0.83 *(0.93)*	— *(0.30)*	0.67
	3	0.85 *(0.63)*	1.18 *(0.17)*	0.83 *(0.81)*	0.88 *(0.07)*	1.44 *(0.06)*	— *(0.31)*	0.71
	4	1.01 *(0.35)*	1.15 *(0.01)*	0.87 *(0.90)*	0.94 *(0.59)*	1.40 *(0.02)*	— *(0.27)*	0.94
UNEMP	0	2.47 *(0.02)*	1.76 *(0.06)*	1.61 *(0.00)*	1.18 *(0.06)*	0.84 *(0.99)*	1.00 *(0.14)*	3.45
	1	2.37 *(0.01)*	1.40 *(0.16)*	1.77 *(0.06)*	1.24 *(0.10)*	1.15 *(0.46)*	— *(0.90)*	1.98
	2	1.30 *(0.25)*	1.18 *(0.33)*	1.30 *(0.26)*	1.16 *(0.47)*	1.38 *(0.11)*	— *(0.98)*	1.31
	3	1.04 *(0.49)*	1.39 *(0.08)*	1.49 *(0.00)*	1.22 *(0.02)*	1.19 *(0.03)*	— *(0.72)*	0.97
	4	0.93 *(0.99)*	1.01 *(0.40)*	1.30 *(0.13)*	1.09 *(0.41)*	1.01 *(0.55)*	— *(0.61)*	1.00

Forecast evaluation from $T_0 + h$ to T_1 where T_0 = 2000Q4, T_1 = 2007Q3, and $h = 0, \ldots, 4$ ($28 - h$ quarterly evaluation points). All combination methods use a rolling training sample of 40 observations.[a] Numbers in parantheses are p-values of the Superior Predictive Ability (SPA) test by Hansen (2005) with corresponding method used as benchmark; we use a mean block length of 8 observations and 10.000 replications in the bootstrap implementation. All other numbers are MSPEs, relative to the mean SPF prediction (*EW*).

[a] Except for the variable TBOND which is only available since 1992Q1 and thus has a smaller training window sample for the first evaluation points.

Curriculum Vitæ

Simon D. Knaus from Hemberg (St. Gallen)

Education

2009 – 2013	PhD in Economics and Finance, University of St. Gallen, Switzerland
2005 – 2007	MSc in Mathematics, University of Fribourg, Switzerland
2001 – 2004	BSc in Mathematics, University of Fribourg, Switzerland

Working Experience

2013 – date	Consultant with c-alm AG, St. Gallen, Switzerland
2008 – 2008	Consultant with McKinsey & Co., Inc., Zürich, Switzerland
2006 – 2008	Treasury IT engineer with Postfinance, Berne, Switzerland
2004 – 2005	Intern with McKinsey & Co., Inc., Zürich, Switzerland
2000 – 2001	Software engineer with Viviance AG, St. Gallen, Switzerland

Conference Presentations

2012	Int. Conference on Computational Statistics, Limassol, Cyprus R/Rmetrics, Leissingen, Switzerland
2011	Academy of Behavioral Finance & Economics, Los Angeles, USA CESifo Conference, Munich, Germany R/Rmetrics, Leissingen, Switzerland
2010	Int. ERCIM Conference Computing & Statistics, 2010, London, UK