Three Essays on Volatility Forecasting and Forecast Evaluation

Inauguraldissertation

zur Erlangung des akademischen Grades Doctor Rerum Politicarum

an der Fakultät für Wirtschafts- und Sozialwissenschaften der
Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Onno Kleen

Heidelberg, Juni 2020

Acknowledgments

During the last five years, I have greatly benefited from interacting with lots of people. I would like to thank all of them, yet some people deserve special mentioning:

Christian, thank you for being my supervisor and for your continuous support from the very beginning. You always took your time for discussing research projects, the lengthy publication process of our first article, and you gave great career advice. I am very grateful that you sent me to a SoFiE conference in Aarhus right at the start of my dissertation. You were always supportive of me presenting my research at conferences and workshops which helped me tremendously to connect with researchers outside of our department. Additionally, the long-term employment security you and Switgard Feuerstein offered me was a great privilege. Thank you for all of that and I very much hope that we will start new research projects together in the future.

Over time I had and continue to have great colleagues in our Macro-Econometrics group who always gave valuable feedback—I very much enjoyed our numerous discussions over lunch. Thank you, Alexander, Christopher, David, Fabian, Jonas, Karin, Lora, Marina, Matthias, Michael, Thomas, and Zeno. Likewise, I benefited tremendously from discussing my research with all members of the HKMetrics network and with my host at Duke University, Andrew.

Melanie, thank you for being a member of my thesis committee in addition to co-founding the HKMetrics network with its great PhD workshops.

Finally, all these years since we left Ostfriesland for Heidelberg would have been less exciting without you, Sarah. Thank you for your unconditional support and I very much look forward for many years to come.

Contents

Introduction

Economic forecasting

Forecasts are ubiquitous in economics and finance as agents make decisions based on uncertainty of future outcomes (Elliott and Timmermann, 2008). For example, the expectations of a household's earnings guide long-term investments such as buying residential property. Similarly, a company's investment decision depends on expected future sales and interest rates. But also public decisions like building new schools or nursing homes are aligned with anticipated demographic changes.

As forecasts are so prevalent for economic decision making, many ways have been thought of to create these forecasts. One approach is to use time series models that are estimated on past data; for example, gross domestic product (GDP) and inflation forecasts by vector autoregression models introduced by Sims (1980) or volatility forecasts for risk management and portfolio allocation by generalized autoregressive conditional heteroskedasticity (GARCH) models (Bollerslev, 1986). These models are typically used to generate *point forecasts.* Most of these forecast procedures target the conditional mean, but there are also notable exceptions such as value-at-risk—a quantile point forecast in the left tail of asset or portfolio return distributions. A second approach is to conduct surveys among experts or consumers. One example is the Survey of Professional Forecasters conducted by the Federal Reserve Bank of Philadelphia in which respondents forecast a rich set of variables.

A considerable drawback of point forecasts is that they do not convey any sense of the expected uncertainty attached to the prediction. However, decisions of consumers, businesses, and investors are not only tied to point predictions of future expected outcomes but to multiple functionals of the subjective uncertainty; for example, downside risk in financial assets. After the financial crisis in 2008, we saw an ominous interest in research on the effect of macroeconomic uncertainty on economic decisions (e.g., Jurado, Ludvigson, and Ng, 2015; Baker, Bloom, and Davis, 2016).

One of the first institutions that communicated its uncertainty about future (and even past) macroeconomic developments to the public is the Bank of England's Monetary Policy Committee who reports *distribution forecasts* of inflation rates since February 1996.[1] Distribution forecasts (sometimes also referred to as probabilistic forecasts) are forecasts that attach probabilities to all possible outcomes. Therefore, they are the most conclusive measure of uncertainty compared to other single-valued risk measures like value-at-risk.

One of the most important questions is: What defines a forecast to be "good"—and what is an evaluation criterion that aligns with the definition of choice. The literature comparing rivaling forecasts is based on loss functions—equivalently called scoring rules—that assign a real value to any pair of forecast and observation. Based on decision theory, these loss functions are typically assumed to be consistent in the case of point forecasts or assumed to be proper in the case of distribution forecasts (Gneiting, 2011; Gneiting and Raftery, 2007) as an incentive for stating honest beliefs about future outcomes. Popular loss functions are the squared error for point predictions targeting the conditional mean or the negative log-likelihood (Good, 1952) for distribution forecasts.

However, in economic forecasting a prevalent problem is that true predictands are not necessarily observable but are often measured with error. Two prominent examples are macroeconomic data being revised through time or stock market volatility. Fortunately, for mean point predictions it can be shown that some loss functions imply the same expected forecast ranking even if unbiased proxy observations are used for evaluating the forecasts (Patton, 2011). For distribution forecasts, however, there is no such result and we address this shortcoming by examining the sensitivity of loss functions for distribution forecasts in the presence of observational error.

Outline of the thesis

The three main chapters of this dissertation are self-contained research articles that can be read independently from each other. They all focus on forecasting with financial and macroeconomic data. The analyses in Chapter 1 and 2 are joint works with Christian Conrad. Both focus on forecasting volatility for financial markets. In Chapter 1, we address aggregate stock market volatility and in Chapter 2 stock-specific volatility for investment decisions. The first of the two has been published under the title "Two are better than one: Volatility forecasting using multiplicative component GARCH-MIDAS models" in the *Journal of Applied Econometrics*

[1] https://www.bankofengland.co.uk/inflation-report/inflation-reports

(Conrad and Kleen, 2020). Chapter 3 is single-authored and, in contrast to the other two chapters, focuses on the evaluation of distribution forecasts. The outline for the separate chapters is as follows.

Chapter 1: Two are better than one: Volatility forecasting using multiplicative component GARCH-MIDAS models

In Chapter 1, we examine the properties and forecast performance of multiplicative volatility specifications that belong to the class of generalized autoregressive conditional heteroskedasticity mixed-data sampling (GARCH-MIDAS) models suggested in Engle, Ghysels, and Sohn (2013). The main idea of these models is to decompose volatility into a short-term GARCH component and a long-term component that is driven by an explanatory variable. The contribution of this chapter to the recent strand of literature is twofold.

In the first part of Chapter 1, we analyze several statistical properties of the GARCH-MIDAS model, namely the kurtosis of returns, the autocorrelation function of squared returns, and the R^2 of a Mincer-Zarnowitz regression. We then evaluate the quasi-maximum likelihood estimate (QMLE) and forecast performance of these models in a Monte-Carlo simulation.

Our main theoretical findings are described as follows. In the GARCH-MIDAS model, the kurtosis of the returns is always larger than the kurtosis of the returns in the nested GJR-GARCH (Glosten, Jagannathan, and Runkle, 1993) component. If the long-term component is sufficiently persistent, the autocorrelation function (ACF) of both the squared returns and the latent conditional variance process is more persistent than the corresponding ACFs in the nested GJR-GARCH. Both findings indicate that a multiplicative component structure in the volatility of stock returns can explain the common failure of simple one-component GARCH models to adequately capture the stylized facts of returns and realized variances. Further, we derive an upper bound for the population R^2 in the k-step-ahead Mincer and Zarnowitz (1969, henceforth MZ) regression of the squared return on the volatility forecast. We show that the population R^2 decreases monotonically in the forecast horizon but increases monotonically in the variability of the long-term component. The latter feature leads to the unpleasant property that the goodness-of-fit is particularly high in situations in which the squared error loss is also high. Clearly, this finding questions the usefulness of the MZ R^2 for comparing forecast accuracy across volatility regimes.

In a Monto-Carlo simulation, we evaluate the QMLE of GARCH-MIDAS models. The QMLE is shown to be unbiased and the asymptotic standard errors based on Wang and Ghysels (2015) are shown to be valid even in the presence of exogenous explanatory variables. Another question that had not been addressed so far is the effect of measurement error in the explanatory variable or a misspecification of the lag structure. Based on our simulation study, we conclude that these have only minor effects. The last result based on our simulations is that even though eliciting the long-term component is a relatively difficult estimation problem, the GARCH-MIDAS model beats two competitor models, the Markov-Switching GARCH (MS-GARCH) by Haas, Mittnik, and Paolella (2004) and the nested GARCH if correctly specified and—at least in most setting—even if its misspecified. The argument is based on an out-of-sample (OOS) forecast evaluation employing the QLIKE loss which is minimized by the true conditional mean forecast, robust to the measurement error in realized volatility, and less sensitive to outliers compared to the squared error (Patton, 2011).

In the second part of Chapter 1, we conduct an OOS forecast performance study targeted at the volatility of the S&P 500 index, in which we compare the GARCH-MIDAS with a wide range of competitor models: the heterogeneous autoregression (HAR) of Corsi (2009), the realized GARCH of Hansen, Huang, and Shek (2012), the high-frequency-based volatility (HEAVY) of Shephard and Sheppard (2010), and the MS-GARCH. For a realistic evaluation of the GARCH-MIDAS models' ability to describe the behavior of long-term financial volatility we make use of real-time data of United States (US) macroeconomic and financial conditions to avoid a look-ahead bias due to publication lag of macroeconomic data. The model evaluation is carried out by constructing model confidence sets (MCS) (Hansen, Lunde, and Nason, 2011) that allow a joint forecast evaluation of more than two models. Our results are that at forecast horizons of two weeks and one month, the MCS consists of the Realized GARCH, the HAR, and GARCH-MIDAS models based on the Chicago Board of Exchange Volatility Index (VIX). At longer forecast horizons of two and three months ahead, only GARCH-MIDAS models are included in the MCS. As in previous studies, the GARCH-MIDAS based on housing starts performs particularly well.

Last, in the course of writing this chapter we developed R packages published on the Comprehensive R Archive Network for downloading real-time data from the ALFRED database of the Federal Reserve Bank of St. Louis (Kleen, 2017) and for forecasting using GARCH-MIDAS models (Kleen, 2018).

Chapter 2: Volatility forecasting for low-volatility investing

In Chapter 2, we examine whether recent advances in volatility forecasting are beneficial for implementing low-volatility portfolios. Low-risk strategies such as betting-against-beta (Frazzini and Pedersen, 2014), low-volatility portfolios (Blitz and van Vliet, 2007), and volatility-managed portfolios (Moreira and Muir, 2017) have become increasingly popular. In practice, these strategies are often based on a rather simple volatility proxy; for example, the sample standard deviation of daily returns over the previous year which is our leading benchmark strategy. Based on the ranking of these volatility proxies, the investor picks, for example, the bottom quintile of stocks to invest in. However, this simple approach is at odds with advances in financial econometrics if the low-volatility classification problem is looked at from a forecasting perspective.

We follow the literature on estimating volatility from intraday return data and measure monthly volatility by realized variances (Andersen et al., 2003). Our first observation is that the infeasible ex-post optimal sorting based on realized variances—which we will refer to as the oracle portfolio—earns higher returns than the one based on daily data. The question is now whether the superior forecast performance of state-of-the-art volatility models in terms of forecast errors translates into superior rankings in real time. We examine this by employing a wide range of Riskmetrics, GARCH-, HAR-, and MIDAS-type models for all real-time constituents of the S&P 500 in each month and use these models to forecast next-month's volatility in between 2002–2018. The simple proxies used in the industry can also be thought of being a (naive) benchmark model or forecast.

In the evaluation of our forecasts, we take two different points of view. First, we aggregate forecast losses per stock over time and compare how often the aggregated losses of our time series models are lower than the one of the benchmark. In line with the literature, we find the HAR-type models to be dominant. Unfortunately, this approach for ranking models on a stock-by-stock basis is practically infeasible in our application due to data restrictions and the time-varying stock universe. Second, we evaluate our models based on their cross-sectional forecast performance which—due to the large cross-section—can be used for model selection in real time. Overall, the HAR models remain dominant but less so; for example, the Realized GARCH becomes the best-performing model in 21% (70%) of months when measured by the squared error (QLIKE) loss function. Combining the models based on cross-sectional losses leads to further improvements in forecast performance for all four evaluation criteria employed.

In a next step, we derive the forecast-implied volatility ranking for all models and loss-based combination forecasts. We compare the resulting low-volatility portfolios with the infeasible oracle portfolio and find that the benchmark portfolio based on the empirical standard deviation of last year's daily returns performs worse in mimicking the oracle portfolio than our approach. The similarity of one of our portfolios to the oracle portfolio is measured through the average number of stocks that are included in both. Likewise, the cross-sectional average volatility inside our portfolios is typically lower.

Even though we improve upon the benchmark strategy in terms of similarity to the oracle portfolio, we do not find significant differences in terms of returns between our strategies and the benchmark portfolio. Some of the best models in terms of forecast performance have higher returns but only before transaction costs are taken into account. We explain this finding by observing that the turnover of our benchmark model is by far the smallest among all strategies considered.

Chapter 3: Measurement error sensitivity of loss functions for distribution forecasts

In Chapter 3, we analyze the sensitivity of distribution forecast evaluation in settings in which the predictand is observed with measurement error or simply measured on different scales. Our result is that the continuous ranked probability score (CRPS) (Matheson and Winkler, 1976) is less sensitive to observational error than the log(-likelihood) score while also being robust to rescaling the data.

In the first part of the theoretical section, we focus on forecast comparison with linearly rescaled data. Here, we see that all commonly used scoring rules imply a robust forecast ranking; that is, even after rescaling the ranking is preserved in expectation. Even though this seems to be a condition every evaluation criteria in economics should satisfy, we show that simple linear combinations of proper scoring rules do not fulfill this criterion. Our results are obtained by introducing the notion of scaling-invariance for loss functions, which is a slightly more general definition of homogeneity than the one used in Patton (2011) for point forecasts.

The second part of the theoretical section is concerned with measurement error in the observations. The expected loss with respect to proper scoring rules as in Gneiting and Raftery (2007) is minimized when forecasters state the true conditional distribution of the observation as a forecast. As a consequence, if we can not observe the predictand directly but only the predictand subject to an error component, proper scoring rules favor forecasts of the observations and not necessarily of the true predictand. For example, in the case of additive measurement error one would prefer distribution forecasts that have a larger variance than the conditional distribution of the true predictand alone.

One approach for addressing this misalignment is to calculate error-corrected scoring rules (Ferro, 2017; Naveau and Bessac, 2018). The idea is to examine the difference in expected loss employing the noisy proxy and the expected loss given the true predictand. However, this approach is tied to knowing the true predictand's distribution, the underlying error distribution, and the specific forecast distribution at hand.

We thus take an alternative approach and examine whether some proper scoring rules are less sensitive to classes of error distributions than others even though both imply biased forecast rankings in expectation. Following the theory of robust estimators by Hampel (1968, 1971), we define a loss function to be gross-error insensitive if the expected absolute deviation in losses with respect to a class of error distributions from the true outcome is bounded across the outcome space. In contrast to specific error-corrected scoring rules, gross-error sensitivity is defined with respect to *classes* of forecast and error distributions. The quadratic score, which is closely related to the log score, and the CRPS turn out to be gross-error insensitive but the log score is not. Our results are in line with the literature discussing the robustness of different scoring rules for M-estimation (Basu et al., 1998; Kanamori and Fujisawa, 2015; Dawid, Musio, and Ventura, 2016; Ovcharov, 2017).

Our theoretical results are illustrated by a simulation study and an empirical application. The main focus is here to review whether the insensitivity of expected losses transfers to less sensitive test statistics of equal predictive ability. In the simulation, the data-generating process is aligned with US GDP growth data and the corresponding revision errors. In the empirical application we forecast daily volatilities of 28 Dow Jones Industrial Average constituents. The comparison between log score and CRPS shows that the latter always leads to a smaller bias in the test statistics; in this case, gross-error insensitivity translates into more stable test outcomes across different measurement errors. However, the favorable result for the CRPS does not hold for every gross-error insensitive loss function. The simulation and empirical illustration show that the quadratic score is only insensitive to infrequent but possible large measurement error.

1 Two are better than one: Volatility forecasting using multiplicative component GARCH-MIDAS models

Abstract

We examine the properties and forecast performance of multiplicative volatility specifications that belong to the class of generalized autoregressive conditional heteroskedasticity mixed-frequency data sampling (GARCH-MIDAS) models suggested in Engle, Ghysels, and Sohn (2013). In those models volatility is decomposed into a short-term GARCH component and a long-term component that is driven by an explanatory variable. We derive the kurtosis of returns, the autocorrelation function of squared returns, and the R^2 of a Mincer-Zarnowitz regression and evaluate the QMLE and forecast performance of these models in a Monte-Carlo simulation. For S&P 500 data, we compare the forecast performance of GARCH-MIDAS models with a wide range of competitor models such as HAR (heterogeneous autoregression), realized GARCH, HEAVY (high-frequency-based volatility) and Markov-switching GARCH. Our results show that the GARCH-MIDAS based on housing starts as an explanatory variable significantly outperforms all competitor models at forecast horizons of 2 and 3 months ahead.

1.1 Introduction

The idea of modeling volatility as consisting of multiple components has a long tradition in financial econometrics (e.g., Ding and Granger, 1996; Engle and Lee, 1999). Early models typically featured additive volatility components and did not allow for explanatory variables in the conditional variance. More recently, the focus has shifted to multiplicative component models (e.g., Engle and Rangel, 2008; Engle, Ghysels, and Sohn, 2013; Amado and Teräsvirta, 2013, 2017; Han and Kristensen, 2015). In particular, the class of generalized autoregressive conditional heteroskedasticity mixed-frequency data sampling (GARCH-MIDAS) models proposed in Engle, Ghysels, and Sohn (2013) has been proven to be useful for analyzing the link between financial volatility and the macroeconomic environment (Asgharian, Hou, and Javed, 2013; Conrad and Loch, 2015; Dorion, 2016). In the GARCH-MIDAS model, a unit-variance GARCH component fluctuates around a time-varying long-term component that is a function of (macroeconomic or financial) explanatory variables. By allowing for a mixed-frequency setting, this approach bridges the gap between daily stock returns and low-frequency (e.g., monthly, quarterly) explanatory variables. For further applications of GARCH-MIDAS-type models see, for example, Conrad, Loch, and Rittler (2014), Opschoor, van Dijk, and van der Wel (2014), Dominicy and Vander Elst (2015), Lindblad (2017), Amendola, Candila, and Scognamillo (2017), Pan et al. (2017), Conrad, Custovic, and Ghysels (2018), and Borup and Jakobsen (2019). For a recent survey on multiplicative component models see Amado, Silvennoinen, and Teräsvirta (2019). Throughout this paper, the GARCH-MIDAS model will be our leading example for a multiplicative component GARCH (M-GARCH) model. However, we will also discuss how the class of M-GARCH models nests other specifications such as the Markov-Switching GARCH (MS-GARCH) of Haas, Mittnik, and Paolella (2004), the Spline-GARCH of Engle and Rangel (2008), and the Multiplicative Time-Varying GARCH (MTV-GARCH) of Amado and Teräsvirta (2008).

Our contribution to this recent strand of literature is twofold. In the first part of this chapter, we analyze several statistical properties of the GARCH-MIDAS model that have not received much attention so far. In the second part of the chapter, we compare the out-of-sample (OOS) forecast performance of the GARCH-MIDAS with the performance of various competitor models such as the heterogeneous autoregression (HAR) of Corsi (2009), the realized GARCH of Hansen, Huang, and Shek (2012), the high-frequency-based volatility (HEAVY) of Shephard and Sheppard (2010), and the MS-GARCH.

Our main theoretical findings can be summarized as follows. In the GARCH-MIDAS model, the kurtosis of the returns is always bigger than the kurtosis of the returns in the nested GJR-GARCH (Glosten, Jagannathan, and Runkle, 1993) component. If the long-term component is sufficiently persistent, the autocorrelation function (ACF) of the squared returns as well as the ACF of the conditional variances is more persistent than the corresponding ACFs in the nested GJR-GARCH. Both findings suggest a multiplicative component structure in the volatility of stock returns as a potential explanation for the common failure of simple one-component GARCH models to adequately capture the stylized facts of returns and realized variances. It should also be noted that our results are remarkably similar to recent findings in Han (2015) on GARCH-X models, even though Han (2015) considers models with an additive explanatory variable in the conditional variance and focuses on the asymptotic limit of the sample kurtosis and the sample ACF. Further, we derive an upper bound for the population R^2 in the k-step-ahead Mincer and Zarnowitz (1969) regression (henceforth MZ regression) of the squared return on the volatility forecast. We show that the population R^2 decreases monotonically in the forecast horizon but increases monotonically in the variability of the long-term component. The latter feature leads to the unpleasant property that the goodness-of-fit is particularly high in situations in which the squared error loss is also high. Clearly, this finding questions the usefulness of the MZ R^2 for comparing forecast accuracy across volatility regimes. In this context, we derive an explicit expression for the one-step-ahead R^2 of the GARCH-MIDAS specification and obtain the results from Andersen and Bollerslev (1998) for the simple GARCH(1,1) as a special case.

Empirically, we first evaluate the quasi-maximum likelihood estimator (QMLE) of GARCH-MIDAS models by means of a Monte-Carlo simulation. We show that the QMLE is unbiased and that the asymptotic standard errors based on Wang and Ghysels (2015) are valid in the presence of exogenous explanatory variables. Further, we show that measurement error in the explanatory variable or a misspecification of the lag structure has only minor effects. We also confirm our theoretical result that the R^2 of a MZ regression is highest in regimes with high volatility although in those regimes forecast performance is the worst. Following the arguments put forth in Patton and Sheppard (2009) and Patton (2011), we use the QLIKE to evaluate the OOS forecast performance of the GARCH-MIDAS model against the MS-GARCH and the nested GARCH. We find that the correctly specified and, in most settings, even the misspecified GARCH-MIDAS models beat the competitor models.

Finally, we apply the GARCH-MIDAS model to a long time series of S&P 500 returns combined with data on US macroeconomic and financial conditions. We consider GARCH-MIDAS models with one or two explanatory variables and, for the OOS forecast evaluation, estimate all models on a rolling window using the appropriate real-time vintage data. Because macroeconomic time series are revised substantially after the first release, we avoid a look-ahead bias by using real-time data. In the OOS forecast evaluation, we compare the GARCH-MIDAS with eight competitor models: Among those competitor models are the Realized GARCH, the HEAVY, the MS-GARCH, and HAR models with and without leverage. We evaluate all models jointly by constructing model confidence sets (MCS) as introduced in Hansen, Lunde, and Nason (2011). For forecast horizons of two weeks and one month, the MCS consists of the Realized GARCH, the HAR, and GARCH-MIDAS models with the Cboe Volatility Index (VIX) (or the VIX combined with another explanatory variable). That is, at these forecast horizons the GARCH-MIDAS is on par with those models but beats the HEAVY as well as MS-GARCH models. At longer forecast horizons of two and three months ahead, only GARCH-MIDAS models are included in the MCS. At both horizons the GARCH-MIDAS based on housing starts achieves the lowest QLIKE. This finding is remarkable because our OOS period begins in 2010 and hence does not include the financial crisis and the collapse of the housing bubble.

To facilitate the replication of our results, we provide R packages for downloading real-time data from the ALFRED database of the Federal Reserve Bank of St. Louis (Kleen, 2017), as well as for estimating GARCH-MIDAS models (Kleen, 2018).[1]

This chapter is organized as follows: In Section 1.2, the M-GARCH model and our theoretical results are presented. In Section 1.3, we perform a simulation study and, in Section 1.4, we apply the GARCH-MIDAS model to S&P 500 return data. The conclusion follows in Section 1.5. All appendices of this chapter can be found in Section 1.6. The proofs are contained in Appendix 1.6.1. Additional material can be found in Appendices 1.6.2–1.6.8.

1.2 The multiplicative component GARCH model

In this section, the M-GARCH model is introduced and its theoretical properties are derived. In particular, we show that the M-GARCH model inherits certain time series properties that are in line with stylized facts typically observed for financial return data but cannot be captured by simple GARCH models.

1.2.1 Model specification

We denote daily log-returns by $r_{i,t}$, whereby the index $t = 1, \ldots, T$ refers to a certain period (e.g. a week or a month) and the index $i = 1, \ldots, I_t$ to days within that period. For simplicity, we model the returns as $r_{i,t} = \mu + \varepsilon_{i,t}$.[2] The M-GARCH model assumes that the scaled (demeaned) returns can be written as

$$\frac{\varepsilon_{i,t}}{\sqrt{\tau_t}} = \sqrt{g_{i,t}} Z_{i,t}, \tag{1.1}$$

[1] The packages are available at: *https://cran.r-project.org/package=alfred* and *https://cran.r-project.org/package=mfGARCH*.

[2] It would be straightforward to allow for richer dynamics in the conditional mean. However, for daily return data a constant conditional mean is usually sufficient. For simplicity, in the following we refer to $\varepsilon_{i,t}$ as the (demeaned) return.

where τ_t is specified as a function of a (low-frequency) explanatory variable X_t, $g_{i,t}$ follows a GARCH equation, and $Z_{i,t}$ is an *i.i.d.* innovation process with mean zero and variance one. Let $\mathcal{F}_{i,t}$ denote the information set up to day i in period t and define $\mathcal{F}_t := \mathcal{F}_{I_t,t}$. If τ_t depends on lagged values of X_t only, then

$$\sigma_{i,t}^2 := g_{i,t}\tau_t$$

is the conditional variance of the daily returns; that is, $\sigma_{i,t}^2 = \mathbf{Var}(\varepsilon_{i,t}|\mathcal{F}_{i-1,t})$. We refer to $g_{i,t}$ as the short-term component of volatility and to τ_t as the long-term component of volatility. While $g_{i,t}$ varies daily, τ_t is constant across all days within period t and thus changes at the lower frequency only. The short-term component is intended to describe the well-known day-to-day clustering of volatility and is assumed to follow a mean-reverting unit-variance GJR-GARCH(1,1) process:

$$g_{i,t} = (1 - \alpha - \gamma/2 - \beta) + \left(\alpha + \gamma \mathbb{1}_{\{\varepsilon_{i-1,t}<0\}}\right) \frac{\varepsilon_{i-1,t}^2}{\tau_t} + \beta g_{i-1,t}. \tag{1.2}$$

Remark 1.1. *We use the convention that $\varepsilon_{0,t} = \varepsilon_{I_{t-1},t-1}$ and $g_{0,t} = g_{I_{t-1},t-1}$. Similarly, we can write the long-term component as $\tau_{i,t} = \tau_t$ for $i = 1, \dots, n$ and $\tau_{0,t} = \tau_{I_{t-1},t-1} = \tau_{t-1}$. That is, for $I_t > 1$, τ_t is piecewise constant. If $I_t = 1$, then both components vary at the same frequency. In this case we can write $\varepsilon_{1,t} = \varepsilon_t$, $g_{1,t} = g_t$, $\varepsilon_{0,t} = \varepsilon_{1,t-1} = \varepsilon_{t-1}$, and $g_{0,t} = g_{1,t-1} = g_{t-1}$. Thus, we can drop the index i.*

A characteristic of the two-component M-GARCH model defined in Equation (1.1) is that the scaled returns, $\varepsilon_{i,t}/\sqrt{\tau_t}$, are assumed to follow a GARCH process. Hence, the forcing variable in Equation (1.2) is $\varepsilon_{i-1,t}^2/\tau_t$. This feature distinguishes the two-component M-GARCH specification from standard GARCH models. In those models it is assumed that $\tau_t = 1$ and hence the returns themselves follow a GARCH process. Similarly, additive component GARCH models, such as the model of Engle and Lee (1999), assume that $\tau_t = 1$ and decompose $g_{i,t}$ into two or more GARCH components (with forcing variable $\varepsilon_{i-1,t}^2$). We make the following assumptions regarding the innovation process $Z_{i,t}$ and the parameters of the short-term component.

Assumption 1.1. *Let $Z_{i,t}$ be i.i.d. with $\mathbf{E}[Z_{i,t}] = 0$, $\mathbf{E}[Z_{i,t}^2] = 1$, and $1 < \kappa < \infty$, where $\kappa = \mathbf{E}[Z_{i,t}^4]$.*

Assumption 1.2. *We assume that $\alpha > 0$, $\alpha + \gamma > 0$, $\beta \geq 0$, and $\alpha + \gamma/2 + \beta < 1$. Moreover, the parameters satisfy the condition $(\alpha + \gamma/2)^2\kappa + 2(\alpha + \gamma/2)\beta + \beta^2 < 1$.*

Assumptions 1.1 and 1.2 imply that $\varepsilon_{i,t}/\sqrt{\tau_t} = \sqrt{g_{i,t}}Z_{i,t}$ is a covariance stationary GJR-GARCH(1,1) process. The first- and second-order moments of $g_{i,t}$ are given by $\mathbf{E}[g_{i,t}] = 1$,

$$\mathbf{E}[g_{i,t}^2] = \frac{1 - (\alpha + \gamma/2 + \beta)^2}{1 - (\alpha + \gamma/2)^2\kappa - 2(\alpha + \gamma/2)\beta - \beta^2}, \tag{1.3}$$

and the fourth moment of $\sqrt{g_{i,t}}Z_{i,t}$ is finite. The role of the second component, τ_t, is to describe smooth movements in the conditional variance. In general, we specify τ_t as a measurable, positive-valued function, $f(\cdot)$, of the present and $K \geq 1$ lagged values of an explanatory variable X_t:

$$\tau_t = f(X_t, X_{t-1}, \ldots, X_{t-K}). \tag{1.4}$$

The appropriate choice of the explanatory variable X_t and of the function $f(\cdot)$ is up to the researcher and will depend on the specific application at hand.[3] The explanatory variable can either vary at the daily frequency (i.e., $I_t = 1$) or at a lower frequency (i.e., $I_t > 1$). Thus, the choice of X_t defines the low frequency t. In GARCH-MIDAS-type models τ_t depends on lagged values of X_t only. By explicitly allowing τ_t to depend on X_t in Equation (1.4), we ensure that our setting also covers MS-GARCH models (see Subsection 1.2.2 for details). We make the following assumption about the explanatory variable X_t and the function $f(\cdot)$:

Assumption 1.3. *Let $f(\cdot) > 0$ be a measurable function and X_t be a strictly stationary and ergodic time series with $\mathbf{E}[|X_t|^q] < \infty$, where q is sufficiently large to ensure that $\mathbf{E}[\tau_t^2] < \infty$. X_t is independent of $Z_{i,t-j}$ for all t, i and j.*

Note that Assumption 1.3 implies that τ_t is strictly stationary (Billingsley, 1995, p. 495), covariance stationary, and independent of the 'GARCH part' (i.e. $g_{i,t-j}Z_{i,t-j}^2$) of the model. In empirical applications the function $f(\cdot) > 0$ is often chosen as being linear in the lagged X_t:

$$\tau_t = m + \pi_1 X_{t-1} + \ldots + \pi_K X_{t-K}. \tag{1.5}$$

The linear specification requires $m > 0$ and $\pi_l \geq 0$, for $l = 1, \ldots, K$, and is feasible only if X_t is a nonnegative variable. If X_t can take positive as well as negative values, it is natural to opt

[3] While we focus on multiplicative GARCH models, Han and Park (2014) and Han (2015) analyze the properties of a GARCH-X specification with an explanatory variable that enters additively into the conditional variance equation. See also Francq and Thieu (2019).

for an exponential specification:

$$\tau_t = \exp(m + \pi_1 X_{t-1} + \ldots + \pi_K X_{t-K}). \tag{1.6}$$

The assumption that X_t is independent of $Z_{i,t-j}$ for all t, i, and j might appear to be rather strong. However, without imposing any restrictions on the functional form of $f(\cdot)$, it greatly simplifies the analysis when discussing the statistical properties of M-GARCH models in Subsection 1.2.3. From an empirical perspective, we believe that it is reasonable to assume that a low-frequency explanatory variable X_t—such as monthly industrial production growth—is (close to being) independent of the daily innovations $Z_{i,t-j}$. For daily explanatory variables (e.g., measures of realized volatility) the independence assumption might appear to be restrictive. However, even if there is a dependence between the innovation to the daily returns and the daily explanatory variable, the dependence between τ_t and $Z_{i,t-j}$ is likely to be negligible. This is because τ_t is a rather smooth function that is obtained as a weighted average of many lags of the daily X_t. Indeed, in Section 1.3 and Appendix 1.6.4 we illustrate in simulations that a mild violation of the independence assumption does not affect our main results.

It should also be noted that the same independence assumption has been previously made in related literature on M-GARCH models, see Han and Kristensen (2015). Nevertheless, it clearly imposes a limitation that should be overcome in future work. Two examples in this direction are the estimation of GARCH-MIDAS models employing lagged values of realized variances (Wang and Ghysels, 2015) and testing for an omitted long-term component in one-component GARCH models (Conrad and Schienle, 2020).

Assumptions 1.1, 1.2, and 1.3 imply that the $\varepsilon_{i,t}$ have mean zero, are uncorrelated, and have an unconditional variance given by $\mathbf{Var}(\varepsilon_{i,t}) = \mathbf{E}[\tau_t]$. Moreover, the unconditional variance of the squared returns is well defined: $\mathbf{Var}(\varepsilon_{i,t}^2) = \kappa \mathbf{E}[\tau_t^2]\mathbf{E}[g_{i,t}^2] - \mathbf{E}[\tau_t]^2$. If the long-term component is constant and chosen as $\tau_t = \omega/(1 - \alpha - \gamma/2 - \beta)$, our model reduces to the GJR-GARCH with intercept ω.

A measure that is often used to quantify the relative importance of the long-term component is the following variance ratio (Engle, Ghysels, and Sohn, 2013):

$$VR = \mathbf{Var}(\log(\tau_t))/\mathbf{Var}(\log(\tau_t g_t)), \tag{1.7}$$

where $g_t = \sum_{i=1}^{I_t} g_{i,t}$. The ratio measures how much of the total variation in the (log) conditional variance can be explained by the variation in the (log) long-term component.

1.2.2 Nested and related specifications

We first discuss two models that are directly nested in the M-GARCH setting. The two models are the GARCH-MIDAS of Engle, Ghysels, and Sohn (2013) and (a restricted version of) the MS-GARCH model of Haas, Mittnik, and Paolella (2004). Closely related are the Spline-GARCH of Engle and Rangel (2008) and the MTV-GARCH of Amado and Teräsvirta (2008). For further models that have a multiplicative component structure see Amado, Silvennoinen, and Teräsvirta (2019).

GARCH-MIDAS

In the GARCH-MIDAS model the long-term component is defined as in Equation (1.5) or (2.8.1), whereby the weights π_l are parsimoniously specified via a weighting scheme. The most common choice for the long-term component is based on the exponential specification with $\pi_l = \theta \cdot \varphi_l(w_1, w_2)$. Here, the parameter θ determines the sign of the effect of the lagged X_t on the long-term component and the weights $\varphi_l(w_1, w_2) \geq 0$ are parameterized via the Beta weighting scheme

$$\varphi_l(w_1, w_2) = \frac{(l/(K+1))^{w_1-1} \cdot (1 - l/(K+1))^{w_2-1}}{\sum_{j=1}^{K} (j/(K+1))^{w_1-1} \cdot (1 - j/(K+1))^{w_2-1}}. \tag{1.8}$$

By construction, the weights sum to one; that is $\sum_{l=1}^{K} \varphi_l(w_1, w_2) = 1$. It directly follows that $\mathbf{E}[\tau_{t+1}|\mathcal{F}_t] = \tau_{t+1}$. Engle, Ghysels, and Sohn (2013) use monthly industrial production growth and monthly inflation as explanatory variables, while Conrad and Loch (2015) employ quarterly macroeconomic variables such as gross domestic product (GDP) growth. For further applications of this model see Asgharian, Hou, and Javed (2013), Opschoor, van Dijk, and van der Wel (2014) and Dorion (2016). Wang and Ghysels (2015) consider the special case that $f(\cdot)$ is linear, $I_t = 1$ and $X_t = \sum_{j=0}^{J-1} \varepsilon_{t-j}^2$. That is, X_t is the realized variance based on the last J daily returns. Note that for this specification X_t and Z_t are dependent and, hence, Assumption 1.3 would be violated.

MS-GARCH

In the MS-GARCH model the returns are given by $\varepsilon_t = \tilde{\sigma}_{X_t,t} Z_t$, where $\{X_t\}$ is a Markov chain with finite state space $S = \{1, 2, \dots, s\}$ and transition matrix $\mathbf{P}$ with typical element $p_{i,j} = P(X_t = j | X_{t-1} = i)$. A restricted version of the MS-GARCH model of Haas, Mittnik, and Paolella (2004) is nested in our setting with $I_t = 1$. This is best illustrated in the case of $s = 2$: We assume that the conditional variances in the regimes differ in the intercepts but have the same ARCH and GARCH parameters; for example, $\tilde{\sigma}^2_{k,t} = \omega_k + \alpha\varepsilon^2_{t-1} + \beta\tilde{\sigma}^2_{k,t-1}$, $k \in S$. Defining $\tau_t = ((2 - X_t)\omega_1 + (X_t - 1)\omega_2)/(1 - \alpha - \beta)$, we can rewrite the returns as $\varepsilon_t = \sqrt{\tilde{\sigma}^2_{X_t,t}} Z_t = \sqrt{g_t \tau_t} Z_t$, where $g_t = (1-\alpha-\beta)+(\alpha Z^2_{t-1}+\beta)g_{t-1}$. Thus, the conditional variance has a multiplicative structure. In the following, we will refer to this model as MS-GARCH with time-varying intercept (MS-GARCH-TVI). Stationarity conditions for MS-GARCH models can be found in Haas, Mittnik, and Paolella (2004).

Spline-GARCH and Multiplicative Time-Varying (MTV) GARCH

In both models it is assumed that $I_t = 1$. The Spline-GARCH model specifies the long-term component as a spline function and chooses $X_t = t$. Similarly, in the MTV-GARCH $f(\cdot)$ is specified in terms of logistic transition functions and $X_t = t/T$ is the rescaled time. Thus in both models the long-term component is a deterministic function of time and hence Assumption 1.3 is violated.

1.2.3 Properties of the M-GARCH model

In the following, we derive properties of M-GARCH models for which Assumptions 1.1, 1.2, and 1.3 are satisfied.

Kurtosis and autocorrelation function

Financial returns are often found to be leptokurtic. Hence, a desirable feature of a volatility model is that it generates returns with a kurtosis that is similar to the one empirically observed for financial returns. Under Assumptions 1.1, 1.2, and 1.3, the kurtosis of the returns defined in Equation (1.1) is given by

$$\mathcal{K}^{MG} = \frac{\mathbf{E}[\varepsilon^4_{i,t}]}{(\mathbf{E}[\varepsilon^2_{i,t}])^2} = \frac{\mathbf{E}[\sigma^4_{i,t}]}{(\mathbf{E}[\sigma^2_{i,t}])^2}\kappa > \kappa.$$

Thus, the kurtosis of the M-GARCH process is larger than the kurtosis of the innovation $Z_{i,t}$. This is a well known feature of GARCH-type processes. The following proposition relates the kurtosis $\mathcal{K}^{MG}$ of the M-GARCH to the kurtosis $\mathcal{K}^{GA}$ of the nested GARCH(1,1).

Proposition 1.1. *Under Assumptions 1.1–1.3, the kurtosis $\mathcal{K}^{MG}$ of an M-GARCH process is given by*

$$\mathcal{K}^{MG} = \frac{\mathbf{E}[\tau_t^2]}{\mathbf{E}[\tau_t]^2} \cdot \mathcal{K}^{GA} \geq \mathcal{K}^{GA},$$

where $\mathcal{K}^{GA} = \kappa \cdot \mathbf{E}[g_{i,t}^2]$ is the kurtosis of the nested GARCH process and where the equality holds if and only if τ_t is constant.

Hence, for nonconstant τ_t the kurtosis $\mathcal{K}^{MG}$ is the product of $\mathcal{K}^{GA}$ and the ratio $\mathbf{E}[\tau_t^2]/\mathbf{E}[\tau_t]^2 > 1$. When $\tau_t = \omega/(1-\alpha-\gamma/2-\beta)$ is constant, Proposition 1.1 nests the kurtosis of the GJR-GARCH model. Thus, for volatile long-term components the kurtosis of an M-GARCH process can be much larger than the kurtosis of the nested GARCH model.[4] Specifically, Proposition 1.1 holds for the GARCH-MIDAS and for the MS-GARCH-TVI defined in Section 1.2.2.

The empirical ACFs of volatility proxies such as squared returns or realized variances are known to be very persistent (e.g., Ding, Granger, and Engle, 1993; Andersen et al., 2003). In particular, squared returns are often found to decay more slowly than the exponentially decaying ACF implied by the simple GARCH(1,1) model. In the literature on GARCH models, this is usually interpreted as either evidence for long memory (e.g., Baillie, Bollerslev, and Mikkelsen, 1996), structural breaks (e.g., Hillebrand, 2005), or an omitted persistent covariate (Han and Park, 2014) in the conditional variance.

The following propositions show that the theoretical ACFs of the M-GARCH process have a much slower decay than the ACF of the nested GARCH component if the long-term component is sufficiently persistent. Hence, the multiplicative structure provides an alternative explanation for the empirical observation of highly persistent ACFs of squared returns or realized variances. For Propositions 1.2 and 1.3, we consider the case that both components are varying at the same frequency; that is, the length of the period t is one day ($I_t = 1$).

Proposition 1.2. *If $I_t = 1$ and Assumptions 1.1-1.3 are satisfied, the ACF, $\rho_k^{MG}(\varepsilon^2)$, of the*

[4] Han (2015) obtains a similar result for the sample kurtosis of the returns from a GARCH-X model with a covariate that can either be stationary or nonstationary.

squared returns from an M-GARCH process is given by

$$\rho_k^{MG}(\varepsilon^2) = \mathbf{Corr}(\varepsilon_t^2, \varepsilon_{t-k}^2) = \rho_k^\tau \frac{\mathbf{Var}(\tau_t)}{\mathbf{Var}(\varepsilon_t^2)} + \rho_k^{GA} \frac{\mathbf{Var}(g_t Z_t^2)}{\mathbf{Var}(\varepsilon_t^2)} \left(\rho_k^\tau \mathbf{Var}(\tau_t) + \mathbf{E}[\tau_t]^2\right) \tag{1.9}$$

with $\rho_k^\tau = \mathbf{Corr}(\tau_t, \tau_{t-k})$ *and*

$$\rho_k^{GA} = \mathbf{Corr}(g_t Z_t^2, g_{t-k} Z_{t-k}^2) = (\alpha + \gamma/2 + \beta)^{k-1} \frac{(\alpha+\gamma/2)(1-(\alpha+\gamma/2)\beta - \beta^2)}{1 - 2(\alpha+\gamma/2)\beta - \beta^2}$$

being the ACF of the GJR-GARCH component.[5]

Proposition 1.2 shows that the ACF of the squared returns is given by the sum of two terms: The first term corresponds to the ACF of the long-term component ρ_k^τ times a constant, whereas the second term equals the exponentially decaying ACF of the nested GARCH model ρ_k^{GA} times a term that depends again on ρ_k^τ. Hence, if τ_t is sufficiently persistent, $\rho_k^{MG}(\varepsilon^2)$ will essentially behave as ρ_k^τ for k large.[6] For τ_t being constant, the first term in Equation (1.9) is equal to zero and the second term reduces to the ACF of an asymmetric GARCH(1,1). Also, note that the ratio $\mathbf{Var}(\tau_t)/\mathbf{Var}(\varepsilon_t^2)$ is closely related to the variance ratio defined in Equation (1.7) and measures how much of the variation in the squared returns can be attributed to the variation in the long-term component; that is, it measures the importance of the long-term component.

Haas, Mittnik, and Paolella (2004, p. 503) make a similar observation for the MS-GARCH-TVI model that we discussed in Subsection 1.2.2. For this model, they show that the autocorrelations of the squared returns decay at a rate of $\max\{\alpha+\beta, \varpi\}$, where $\varpi = p_{1,1} + p_{2,2} - 1$ is the degree of persistence due to the Markov effects.[7] If ϖ is close to one—that is, if the long-term component is very persistent—the decay rate of this component dominates the decay of the autocorrelation function.

A standard misspecification test for GARCH models is the Ljung-Box statistic applied to the squared deGARCHed residuals, ε_t^2/g_t. The result in Proposition 1.2 may explain why in empirical applications the null hypothesis of this test is often rejected. In the multiplicative model, the ACF of the squared deGARCHed residuals is given by $\rho_k^\tau \cdot \mathbf{Var}(\tau_t)/(\kappa \mathbf{E}[\tau_t^2] - \mathbf{E}[\tau_t]^2)$, which follows the rate of decay of the long-term component and hence is still persistent. Using

[5] Note that ρ_k^{GA} reduces to the ACF of a (symmetric) GARCH(1,1) when $\gamma = 0$ (Karanasos, 1999).

[6] Again, Han (2015) also obtains a bicomponent structure for the sample ACF of the squared returns from a GARCH-X model with a fractionally integrated covariate. Similarly, Han and Kristensen (2015) show that the empirical ACF in a multiplicative model can display long-memory-type behavior.

[7] Haas, Mittnik, and Paolella (2004) consider a symmetric GARCH. Hence, the persistence in the GARCH component is $\alpha + \beta$.

similar arguments to those in the proof of Proposition 1.2, we can derive the ACF of σ_t^2.

Proposition 1.3. *If $I_t = 1$ and Assumptions 1.1–1.3 are satisfied, the ACF, $\rho_k^{MG}(\sigma^2)$, of σ_t^2 is given by*

$$\rho_k^{MG}(\sigma^2) = \mathbf{Corr}(\sigma_t^2, \sigma_{t-k}^2) = \rho_k^{\tau} \frac{\mathbf{Var}(\tau_t)}{\mathbf{Var}(\sigma_t^2)} + \rho_k^{g} \frac{\mathbf{Var}(g_t)}{\mathbf{Var}(\sigma_t^2)} \left(\rho_k^{\tau} \mathbf{Var}(\tau_t) + \mathbf{E}[\tau_t]^2 \right) \tag{1.10}$$

with ρ_k^{τ} as before and $\rho_k^{g} = \mathbf{Corr}(g_t, g_{t-k}) = (\alpha + \gamma/2 + \beta)^k$ being the ACF of the g_t component.

Again, Assumption 1.3 holds for the GARCH-MIDAS and the MS-GARCH-TVI.

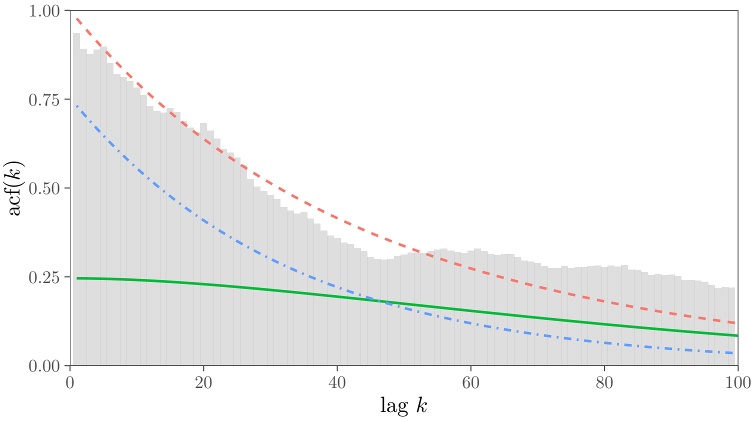
The implications of Proposition 1.3 are depicted in Figure 1.1. The bars in light gray display the empirical ACF of the daily S&P 500 realized variances for the 2000:M1 to 2018:M4 period.[8] The autocorrelations were estimated using the instrumental variables estimator suggested in Hansen and Lunde (2014). We employ their preferred specification, a two-Stage least squares estimator in which lagged realized variances of order four to ten are used as instrumental variables (Hansen and Lunde, 2014, p. 82). By choosing appropriate parameter values for a GARCH-MIDAS process, we obtain an ACF of σ_t^2 (dashed red line) which behaves very similar to the empirical ACF of the realized volatilities. The figure shows that the second term—that is, the ACF of g_t (dot-dashed blue line)—determines the decay behavior of $\rho_k(\sigma^2)^{MG}$ when k is small, while the first term—that is, the ACF of τ_t (solid green line)—dominates when k is large. Finally, it is important to note that although our results on the kurtosis and the ACFs are presented for a GJR-GARCH(1,1) short-term component, they directly extend to a covariance stationary GJR-GARCH(p, q) component.

Forecast evaluation with Mincer-Zarnowitz regression

In empirical applications, the coefficient of determination from a MZ regression is often used as a measure of forecast accuracy. In this section, we will argue against using this measure when comparing forecast performance across volatility regimes. We now exclusively focus on the case of a GARCH-MIDAS. We assume that forecasts are produced at the last day I_t of period t and denote the k-step-ahead volatility forecast by $h_{k,t+1|t}$ with $k \leq I_{t+1}$. The optimal forecast from the GARCH-MIDAS is $h_{k,t+1|t} = \mathbf{E}[\sigma_{k,t+1}^2 | \mathcal{F}_t] = \tau_{t+1} g_{k,t+1|t}$, where $g_{k,t+1|t} = \mathbf{E}[g_{k,t+1} | \mathcal{F}_t] = 1 + (\alpha + \gamma/2 + \beta)^{k-1}(g_{1,t+1|t} - 1)$. When evaluating the volatility forecast, one has to deal with the problem that the true conditional variance, $\sigma_{k,t+1}^2$, is unobservable. Patton (2011) discusses

[8] The underlying data will be described in detail in Subsection 1.4.1.

Figure 1.1: Autocorrelation function of the volatility process in a GARCH-MIDAS model.



Notes: We depict the ACF of the volatility process in a GARCH-MIDAS model (red, dashed) and its components: the first (green, solid) and second term (blue, dot-dashed) in Equation (1.10). The long-term component is defined as in Equation (2.8.1) and Equation (2.5) with $m = -0.1$, $\theta = 0.3$, $w_1 = 1, w_2 = 5$, and $K = 264$. The explanatory variable is given by $X_t = \phi X_{t-1} + \xi_t$, $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\xi^2)$, where $\phi = 0.98$ and $\sigma_\xi^2 = 0.35^2$. The GARCH(1,1) parameters are $\alpha = 0.06$, $\beta = 0.91$, and $\gamma = 0$. Moreover, we set $\kappa = 3$. Bars in light gray display the empirical autocorrelation of S&P 500 daily realized variances in between 2000:M1 and 2018:M4 as measured by Hansen and Lunde (2014). For details see Section 1.4.

the situation in which the forecast evaluation is based on some conditionally unbiased volatility proxy $\hat{\sigma}^2_{k,t+1}$ instead. He defines a loss function $L(\sigma^2_{k,t+1}, h_{k,t+1|t})$ as *robust* if the expected loss ranking of two competing forecasts is preserved when replacing $\sigma^2_{k,t+1}$ by $\hat{\sigma}^2_{k,t+1}$. In the MZ regression $\sigma^2_{k,t+1}$ is often replaced by the conditionally unbiased but noisy proxy $\hat{\sigma}^2_{k,t+1} = \varepsilon^2_{k,t+1}$.[9]

The MZ regression for evaluating the k-step-ahead volatility forecast is given by:

$$\varepsilon^2_{k,t+1} = \delta_0 + \delta_1 h_{k,t+1|t} + \eta_{k,t+1}.$$

We denote the respective coefficient of determination by R^2_k. As shown in Hansen and Lunde (2006), the ranking of competing one-step-ahead volatility forecasts based on the R^2_1 of the MZ regression is robust to using the proxy $\varepsilon^2_{1,t+1}$ instead of the latent conditional variance $\sigma^2_{1,t+1}$ as the dependent variable. For $h_{k,t+1|t} = \tau_{t+1} g_{k,t+1|t}$, the population parameters of the MZ regression are given by $\delta_0 = 0$ and $\delta_1 = 1$ and hence the population R^2_k can be written as:

$$R^2_k = 1 - \frac{\mathbf{Var}(\eta_{k,t+1})}{\mathbf{Var}(\varepsilon^2_{k,t+1})} = 1 - \frac{\mathbf{E}[\mathrm{SE}(\varepsilon^2_{k,t+1}, h_{k,t+1|t})]}{\mathbf{Var}(\varepsilon^2_{k,t})}, \tag{1.11}$$

where we use that the variance of $\eta_{k,t+1}$ equals the expected squared error (SE) loss of the

[9] To illustrate the severeness of the noise, consider an example with $Z_{k,t+1} \sim \mathcal{N}(0,1)$. Then $\varepsilon^2_{k,t+1}$ will either over- or underestimate the true $\sigma^2_{k,t+1}$ by more than 50% with a probability of about 74%.

forecast evaluated against $\varepsilon^2_{k,t+1}$; that is $\mathbf{E}[\mathrm{SE}(\varepsilon^2_{k,t+1}, h_{k,t+1|t})] = \mathbf{E}[(\varepsilon^2_{k,t+1} - h_{k,t+1|t})^2]$. Using that $\mathbf{E}[\varepsilon^2_{k,t+1}|\mathcal{F}_{k-1,t+1}] = \sigma^2_{k,t+1}$, it follows that

$$\mathbf{E}[\mathrm{SE}(\varepsilon^2_{k,t+1}, h_{k,t+1|t})] = \mathbf{E}[\mathrm{SE}(\sigma^2_{k,t+1}, h_{k,t+1|t})] + (\kappa - 1)\mathbf{E}[\sigma^4_{k,t+1}]. \tag{1.12}$$

That is, the expected SE based on the noisy proxy equals the expected SE based on the latent volatility plus a term that depends on the fourth moment, κ, of $Z_{i,t}$ and the expected value of the squared conditional variance. Hence, using a noisy proxy for forecast evaluation can lead to a substantially higher expected SE than the expected SE based on the latent volatility. Patton (2011, p. 248) basically makes the same point by arguing that "although the ranking obtained from a robust loss function will be invariant to noise in the proxy, the actual *level* of expected loss obtained using a proxy will be larger than that which would be obtained when using the true conditional variance."

Using the insight from Equation (1.12) that the expected SE loss based on the noisy proxy is at least $(\kappa - 1)\mathbf{E}[\sigma^4_{k,t}]$, we obtain the following bound:

$$R^2_k \leq 1 - \frac{(\kappa - 1)\mathbf{E}[\sigma^4_{k,t}]}{\kappa\mathbf{E}[\sigma^4_{k,t}] - (\mathbf{E}[\sigma^2_{k,t}])^2} = \frac{1 - (\mathbf{E}[\sigma^2_{k,t}])^2/\mathbf{E}[\sigma^4_{k,t}]}{\kappa - (\mathbf{E}[\sigma^2_{k,t}])^2/\mathbf{E}[\sigma^4_{k,t}]} < \frac{1}{\kappa}. \tag{1.13}$$

The upper bound for R^2_k given by Equation (1.13) nicely illustrates that a low R^2_k is not necessarily evidence for model misspecification but can simply be due to using a noisy volatility proxy. This point has been made before by Andersen and Bollerslev (1998), but for the special case of a one-step-ahead forecast from a GARCH(1,1).[10] Note that the result in Equation (1.13) does not depend on the two-component structure of the model but is true for any conditionally heteroskedastic process.

Next, we derive an explicit expression for the Mincer-Zarnowitz R^2_k of the GARCH-MIDAS model.

Proposition 1.4. *If $\varepsilon^2_{k,t+1}$ follows a GARCH-MIDAS process, Assumptions 1.1–1.3 are satisfied, and $h_{k,t+1|t} = \tau_{t+1}g_{k,t+1|t}$, then the population R^2_k of the MZ regression is given by*

$$R^2_k = \frac{\mathbf{Var}(h_{k,t+1|t})}{\mathbf{Var}(\varepsilon^2_{k,t+1})} = \frac{\mathbf{E}[g^2_{k,t+1|t}]\mathbf{E}[\tau^2_{t+1}] - \mathbf{E}[\tau_{t+1}]^2}{\mathbf{E}[g^2_{k,t+1}]\mathbf{E}[\tau^2_{t+1}]\kappa - \mathbf{E}[\tau_{t+1}]^2}$$

[10] See Andersen, Bollerslev, and Meddahi (2005) for a model-free adjustment procedure for the predictive R^2.

with $\mathbf{E}[g^2_{k,t+1}]$ *as in Equation (1.3) and*

$$\mathbf{E}[g^2_{k,t+1|t}] = 1 + (\alpha + \gamma/2 + \beta)^{2(k-1)}(\mathbf{E}[g^2_{1,t+1}] - 1).$$

We obtain the following two properties:

1. R^2_k *decreases monotonically with increasing forecast horizon* k *and, in the limit, converges*[11] *to* $R^2_\infty = \mathbf{Var}(\tau_{t+1})/\mathbf{Var}(\varepsilon^2_{k,t+1})$.
2. R^2_k *increases monotonically in* $\mathbf{E}[\tau^2_{t+1}]$.

The first property rests on the insight that the forecast of the GARCH component converges to one (as $k \to \infty$) and, hence, the MZ regression reduces to a regression of $\varepsilon^2_{k,t+1}$ on a constant and τ_{t+1}. Thus, the R^2_∞ can be interpreted as the fraction of the total variation in daily returns that can be attributed to the variation in the long-term component. Note that R^2_∞ corresponds to the weight that is attached to the ACF of τ_t in the first term in Equation (1.9).

Second, the result that R^2_k increases when τ_{t+1} gets more volatile implies that for the very same model the R^2_k will be higher in high-volatility regimes (i.e., when the squared error loss is high) than in low-volatility regimes (i.e., when the squared error loss is low). This can be misleading when calculating R^2_k for different regimes. The intuition is best illustrated when looking at one-step-ahead forecasts. Equations (1.11) and (1.12) imply

$$R^2_1 = 1 - \frac{\mathbf{E}[\mathrm{SE}(\varepsilon^2_{1,t+1}, h_{1,t+1|t})]}{\mathbf{Var}(\varepsilon^2_{1,t+1})} = 1 - \frac{(\kappa - 1)\mathbf{E}[g^2_{1,t+1}]\mathbf{E}[\tau^2_{t+1}]}{\mathbf{E}[g^2_{1,t+1}]\mathbf{E}[\tau^2_{t+1}]\kappa - \mathbf{E}[\tau_{t+1}]^2}. \tag{1.14}$$

When $\mathbf{E}[\tau^2_{t+1}]$ is increasing, the unconditional variance of returns rises at a faster rate than the expected squared error and hence the MZ R^2_1 is increasing. We can express R^2_1 directly as a function of the model parameters:

Lemma 1.1. *If* $\varepsilon^2_{k,t+1}$ *follows a GARCH-MIDAS process, Assumptions 1.1–1.3 are satisfied, and* $h_{1,t+1|t} = \tau_{t+1}g_{1,t+1}$, *then the population* R^2_1 *of the MZ regression is given by*

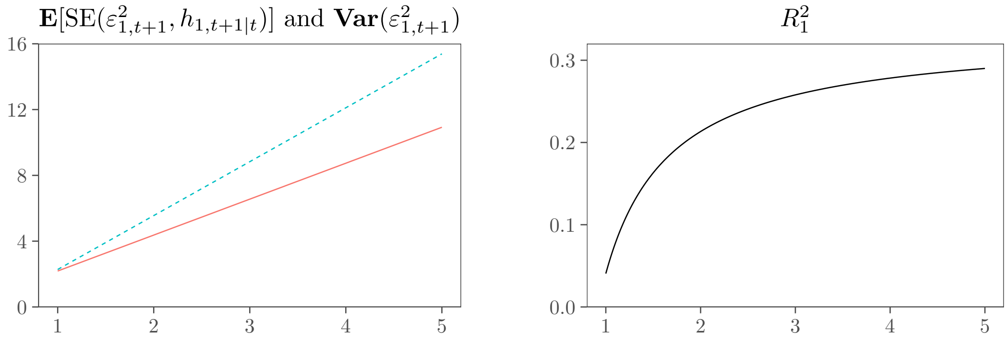
$$R^2_1 = \frac{(1 - (\alpha + \gamma/2 + \beta)^2)\mathbf{E}[\tau^2_{t+1}] - (1 - (\alpha + \gamma/2)^2\kappa - 2(\alpha + \gamma/2)\beta - \beta^2)\mathbf{E}[\tau_{t+1}]^2}{(1 - (\alpha + \gamma/2 + \beta)^2)\mathbf{E}[\tau^2_{\tau+1}]\kappa - (1 - (\alpha + \gamma/2)^2\kappa - 2(\alpha + \gamma/2)\beta - \beta^2)\mathbf{E}[\tau_{t+1}]^2}. \tag{1.15}$$

For τ_{t+1} being constant and $\gamma = 0$, Equation (1.15) is reduced to the expression in Andersen and Bollerslev (1998, p. 892) for the symmetric GARCH(1,1); that is, $R^2_1 = \alpha^2/(1 - 2\alpha\beta - \beta^2)$.

[11] Although by assumption $k \leq I_t$ in our setting, we can think of, for example, a semiannual period and daily volatility forecasts. In this case k can be at most 132 ($= 6 \cdot 22$). For such a large k and under reasonable assumptions on the GARCH parameters, we have $\mathbf{E}[g^2_{132,t+1|t}] \approx 1$.

The effect of an increase in $\mathbf{E}[\tau_{t+1}^2]$ on $\mathbf{E}[\mathrm{SE}(\varepsilon_{1,t+1}^2, h_{1,t+1|t})]$, $\mathbf{Var}(\varepsilon_{1,t+1}^2)$ and R_1^2 is illustrated in Figure 1.2. We set $\mathbf{E}[\tau_{t+1}] = 1$, $\alpha = 0.05, \beta = 0.92, \gamma = 0$, and $\kappa = 3$. As expected, the left panel shows that the expected squared error increases when we move from a low-volatility regime (say $\mathbf{E}[\tau_{t+1}^2] = 2$) to a high-volatility regime (say $\mathbf{E}[\tau_{t+1}^2] = 5$). However, it also shows that the variance of the returns is increasing even faster (as evident from the larger slope coefficient). The right panel of Figure 1.2 shows that this translates into an increase of R_1^2. That is, although the expected squared error increases, the "forecast accuracy" as measured by R_1^2 increases as well. In this regard, the R^2 of a MZ regression should be interpreted as a measure of *relative* forecast accuracy; that is, forecast accuracy is measured relative to the unconditional variance of the process. In contrast, the squared error loss is a measure of *absolute* forecast accuracy. Note that for rather moderate values of $\mathbf{E}[\tau_{t+1}^2]$ the coefficient of determination is already close to its upper bound of 1/3.

Figure 1.2: $\mathbf{E}[\mathrm{SE}(\varepsilon_{1,t+1}^2, h_{1,t+1|t})]$, $\mathbf{Var}(\varepsilon_{1,t+1}^2)$, and MZ R_1^2 as a function of $\mathbf{E}[\tau_{t+1}^2]$.



Notes: The left panel shows $\mathbf{E}[\mathrm{SE}(\varepsilon_{1,t+1}^2, h_{1,t+1|t})]$ (red, solid) and $\mathbf{Var}(\varepsilon_{1,t+1}^2)$ (blue, dashed) as a function of $\mathbf{E}[\tau_{t+1}^2]$ (see Equation (1.14)). The right panel depicts the corresponding population Mincer-Zarnowitz R_1^2 as a function of $\mathbf{E}[\tau_{t+1}^2]$. We set $\mathbf{E}[\tau_{t+1}] = 1$, $\alpha = 0.05, \beta = 0.92, \gamma = 0$, and $\kappa = 3$.

Although the previous results are derived under the assumption that squared daily returns are used as the volatility proxy, it is true that the main insights still hold when using a better volatility proxy. For example, consider the hypothetical case of observing $\sigma_{k,t+1}^2$ ex-post. Then, for $k \to \infty$ we obtain $R_\infty^2 = \mathbf{Var}(\tau_{t+1})/\mathbf{Var}(\sigma_{k,t+1}^2) < 1$. Hence, R_∞^2 would still vary across volatility regimes and increase in the variance of the long-term component. In the simulation in Section 1.3, we will consider the case in which the realized variance is used as a proxy for $\sigma_{k,t+1}^2$.

Finally, we consider cumulative volatility forecasts. The MZ regression for evaluating the cumulative k-day-ahead volatility forecast is given by

$$\widetilde{RV}_{1:k,t+1} = \tilde{\delta}_0 + \tilde{\delta}_1 h_{1:k,t+1|t} + \eta_{1:k,t+1},$$

where the latent variance is proxied by the realized variance $\widetilde{RV}_{1:k,t+1} = \sum_{i=1}^{k} \varepsilon_{i,t+1}^2$ (purely based on daily return data) and $h_{1:k,t+1|t} = \sum_{j=1}^{k} h_{j,t+1|t}$. The corresponding $R_{1:k}^2$ is given by

$$R_{1:k}^2 = \frac{\mathbf{Var}(h_{1:k,t+1|t})}{\mathbf{Var}(\widetilde{RV}_{1:k,t+1})} = \frac{\mathbf{E}[\tau_{t+1}^2]\mathbf{E}[(\sum_{i=1}^{k} g_{i,t+1|t})^2] - k^2\mathbf{E}[\tau_{t+1}]^2}{\mathbf{E}[\tau_{t+1}^2]\mathbf{E}[(\sum_{i=1}^{k} g_{i,t} Z_{i,t}^2)^2] - k^2\mathbf{E}[\tau_{t+1}]^2}.$$

As before, one can show that $R_{1:k}^2$ increases monotonically in $\mathbf{E}[\tau_{t+1}^2]$.

1.2.4 Forecasting long-term volatility

In the empirical application and in the simulation in Section 1.3 we also consider forecasting volatility for horizons that are beyond one low-frequency period. The optimal forecast $h_{k,t+s|t}$ with $s > 1$ is then given by $\mathbf{E}[\tau_{t+s}|\mathcal{F}_t]\mathbf{E}[g_{k,t+s}|\mathcal{F}_t]$. It is straightforward to obtain $g_{k,t+s|t} = \mathbf{E}[g_{k,t+s}|\mathcal{F}_t] = 1 + (\alpha + \gamma/2 + \beta)^{(I_{t+1}+\ldots+I_{t+s-1}+k-1)}(g_{1,t+1|t} - 1)$. Because we do not explicitly model the dynamics of X_t, we are unable to obtain $\mathbf{E}[\tau_{t+s}|\mathcal{F}_t]$. Instead, based on the information set $\mathcal{F}_t$, we forecast τ_{t+s} by τ_{t+1}. Holding the long-term component constant when forecasting is reasonable if τ_t changes smoothly and the forecast horizon is not “too large.” Otherwise, one may use predictions of X_t—for example, survey or time series forecasts—for calculating predictions of τ_t (Conrad and Loch, 2015).

1.3 Simulation

In this section, we mainly focus on M-GARCH models from the GARCH-MIDAS class. Since asymptotic theory for the QMLE is available only for the special case of a GARCH-MIDAS with realized volatility as the explanatory variable (Wang and Ghysels, 2015), we first evaluate the finite-sample performance of the QMLE in a Monte-Carlo simulation. Second, we compare the QMLE of the correctly specified model with the QMLE of misspecified models. We consider misspecification in terms of (i) lag length K, (ii) the explanatory variable being measured with noise, (iii) both, or (iv) omitting the long-component completely. Finally, within the Monte-Carlo simulation we evaluate the OOS forecast performance of the different models listed above and provide empirical support for the theoretical results in Subsection 1.2.3. For each model specification, we perform 2,000 Monte-Carlo replications.

1.3.1 Data generating process

We simulate an intraday version of the two-component GARCH model as

$$\varepsilon_{n,i,t} = \sqrt{g_{i,t}\tau_t} Z_{n,i,t}/\sqrt{N}, \tag{1.16}$$

where the index $n = 1, \ldots, N$ now denotes the intraday frequency. The $Z_{n,i,t}$ are assumed to be i.i.d. and follow either a standard normal or a standardized Student's t distribution with five degrees of freedom. We generate $N = 48$ intraday returns. Hence, by aggregating returns to a daily frequency, $\varepsilon_{i,t} = \sum_{n=1}^{N} \varepsilon_{n,i,t}$, the model in Equation (1.16) is consistent with our daily model.[12] Simulating intraday returns allows us to calculate the daily realized variance, $RV_{i,t} = \sum_{n=1}^{N} \varepsilon_{n,i,t}^2$, as a precise measure of the daily variance. Similarly, we obtain the realized variance over the first k days of month t as $RV_{1:k,t} = \sum_{i=1}^{k} RV_{i,t}$. We simulate data for a period of 40 years of intradaily returns, from which we construct 10,560 daily return and realized variance observations. The parameters of the GARCH-component, $g_{i,t}$, are given by $\alpha = 0.06$, $\beta = 0.91$ and $\gamma = 0$. We consider two alternative specifications of the long-term component:

Monthly τ_t. The first specification assumes a mixed-frequency setting with τ_t fluctuating at a monthly frequency. We assume that each month consists of $I_t = 22$ days. As in Equation (2.8.1), we choose an exponential specification for the long-term component and specify the MIDAS weights according to the Beta weighting scheme in Equation (2.5) with $m = 0.1$, $\theta = 0.3$, $w_1 = 1$, $w_2 = 4$, and $K = 36$. The choice of three years as MIDAS lag length follows Conrad and Loch (2015). Setting $w_2 = 4$ implies a monotonically decaying weighting scheme with weights close to zero for lags greater than two-thirds of K. The explanatory variable X_t is assumed to follow an AR(1) process, $X_t = \phi X_{t-1} + \xi_t$, $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \sigma_\xi^2)$, with $\phi = 0.9$ and $\sigma_\xi^2 = 0.3^2$. When averaged over the 2,000 Monte-Carlo simulations, these parameter values lead to an empirical VR of 18.60%/18.09% for normally/Student's t distributed innovations (recall that the VR was defined in Equation (1.7)).

Daily τ_t. The second specification assumes that both components fluctuate at a daily frequency (i.e. $I_t = 1$). The parameters of the long-term component are chosen as $m = -0.1$,

[12] Alternatively, we simulated the intraday returns using a stochastic volatility model that is consistent with our GARCH-MIDAS setting. The corresponding results, which are very similar to the ones based on the specification in Equation (1.16), are presented in Appendix 1.6.5.

$\theta = 0.3$, $w_1 = 1$, $w_2 = 5$, and $K = 264$. Choosing a lag length of roughly one year is motivated by our empirical results in Section 1.4 when estimating a GARCH-MIDAS model using realized volatility as the explanatory variable. In addition, we choose $\phi = 0.98$ and $\sigma_\xi^2 = 0.2^2$. In the simulations, the former choice leads to an average VR of 32.49%/31.66% for normally/Student's t distributed innovations.

1.3.2 Parameter estimates

Correctly specified models: Bias and asymptotic standard errors

We use the first 20 years of simulated data as the "in-sample" period to obtain QML estimates of the model parameters. Table 1.1 reports the average bias of the QMLE across the 2,000 Monte Carlo simulations. In Panels A/B the innovations $Z_{n,i,t}$ are normally/Student's t distributed. First, we focus on Panel A. In this case the density is correctly specified and the QMLE is the maximum likelihood estimator. Note that for all parameters except w_2 the average bias is close to zero when the conditional variance is correctly specified (i.e., with MIDAS lag length of $K = 36$ (monthly) and $K = 264$ (daily) respectively). For w_2 we clearly observe an upward bias.[13] Based on the 2,000 Monte Carlo replications, we also calculate the empirical standard deviation of the estimated parameters. In Table 1.1 these figures are presented in curly brackets. The numbers in parentheses are the average asymptotic standard errors based on the results in Wang and Ghysels (2015). A comparison of these numbers shows that the asymptotic standard errors are close to the empirical standard deviation of estimated parameters. The only exception is the specification with monthly τ_t where the asymptotic standard errors of w_2 appear to be too big. Nevertheless, the overall performance of the asymptotic standard errors is very satisfying. That is, the Wang and Ghysels (2015) asymptotic standard errors that were derived under the assumption that $X_t = \sum_{j=0}^{J-1} \varepsilon_{t-j}^2$ are applicable more generally.

Misspecified models: Bias

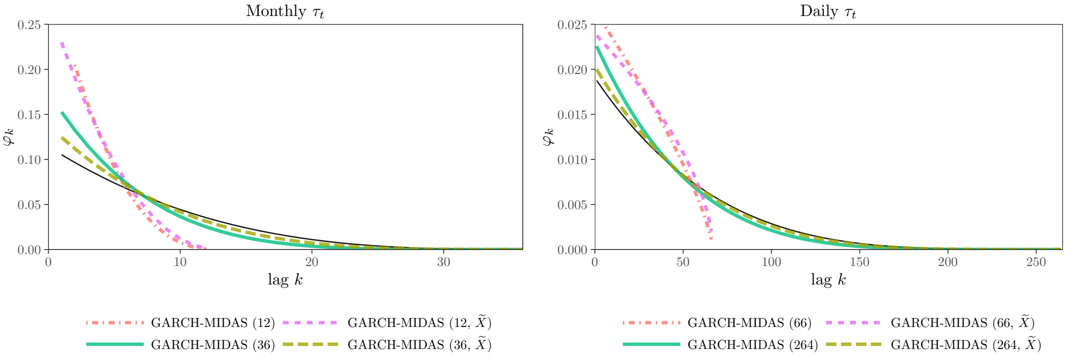
Next, we investigate the effect of model misspecification. First, we consider specifications with a smaller lag length than the true one.[14] Choosing a lag length that is too small ($K = 12$ for

[13] Figure 1.7 in the Appendix compares the histogram of the standardized parameter estimates over the 2,000 Monte Carlo replications with a standard normal distribution. The figure shows that for all parameters except w_2 the empirical distribution of the parameter estimates is very well approximated by the normal distribution.

[14] We do not report results for K being chosen too large as the Beta weighting scheme is flexible enough to downweight uninformative lags to almost zero.

monthly τ_t or $K = 66$ for daily τ_t) does not lead to a bias in the parameter estimates—with the exception of w_2. Now the QMLE of w_2 is downwardly biased. As the estimated weighting schemes in Figure 1.3 show, the downward bias in w_2 translates into biased weighting schemes. Second, we consider the case of observing the explanatory variable X_t with measurement error. This is a reasonable scenario because in practice the true X_t is either unknown to or unobservable for the researcher who will base her analysis on a reasonable proxy. We denote the proxy by $\tilde{X}_t$ and specify it as X_t plus conditionally heteroscedastic noise. In the case of monthly τ_t the noise is given by $\mathcal{N}(0, 0.2 + 0.8|X_t|)$ and in the case of daily τ_t by $\mathcal{N}(0, 0.5 + 0.8|X_t|)$. The average correlation between X_t and $\tilde{X}_t$ is 68.79%/62.71% for monthly/daily τ_t. As before, only the QML estimates of w_2 appear to be biased when X_t is replaced with $\tilde{X}_t$. Last, we estimate a misspecified one-component GARCH model that is obtained when restricting τ_t to be constant. Despite the omitted long-term component, the parameter estimates of α and β are essentially unbiased.

Figure 1.3: Weighting schemes implied by mean parameter estimates.



Notes: Estimated Beta weighting schemes (see Equation (2.5)) as implied by the mean parameter estimates reported in Table 1.1. The green (solid) line corresponds to the case of a correctly specified model whereas the red (dot-dashed) line corresponds to a model with K being too small. With the brown (long dashed) and purple (short dashed) line, the corresponding cases of a GARCH-MIDAS with measurement error are reported. The black line shows the true weighting scheme.

Note that the numbers in Panel B of Table Table 1.1 are very similar to the ones in Panel A. When replacing the normally distributed innovations with Student's t distributed innovations, the density in the maximum likelihood estimation is misspecified and the estimator is truly QMLE. Nevertheless, this change hardly affects our findings. The only notable difference can be seen in the last column of Table 1.1 which shows the average excess kurtosis of the fitted standardized residuals. Those residuals are given by $\varepsilon_{i,t}/\sqrt{\hat{\tau}_t \hat{g}_{i,t}}$ for the GARCH-MIDAS models and by $\varepsilon_{i,t}/\sqrt{\hat{g}_{i,t}}$ for the GARCH model. While the excess kurtosis is essentially zero in Panel

A, in Panel B there is still excess kurtosis, reflecting the fact that the innovations are Student's t distributed.

1.3.3 Forecast evaluation

Next, we evaluate the forecast performance of the different specifications. Based on the in-sample parameter estimates, we construct OOS volatility forecasts for the remaining 20 years. Keeping the parameter estimates fixed is usually referred to as a "fixed (forecasting) scheme."[15] The forecast performance of the different models will be evaluated over the 2,000 Monte-Carlo replications.

We compare the forecast performance of the correctly specified GARCH-MIDAS with all the misspecified models presented in Table 1.1. In addition, we consider the two-state MS-GARCH-TVI model that was introduced in Subsection 1.2.2.[16]

MZ regression

We first present the outcomes of MZ regressions. Figure 1.4 shows the R_k^2 of MZ regressions for volatility forecasts, $h_{k,t+1|t}$, with $k = 1, \ldots, 22$ (i.e., for up to one month ahead). Forecast evaluation is based on the noisy proxy $\varepsilon_{k,t+1}^2$, whereby the data generating process is the GARCH-MIDAS with monthly τ_t and normally distributed innovations. The forecasts are generated from the correctly specified GARCH-MIDAS model. We present the R_k^2 for the full OOS period as well as for three different volatility regimes: *low*, *normal* and *high*. Volatility regimes are defined as follows: We consider the empirical distribution of daily realized variances during the OOS period. A forecast falls into the low/normal/high volatility regime if the level of the realized variance on the day the forecast has been issued is below the 25% quantile, between the 25% and 75% quantile, or above the 75% quantile of the empirical distribution. In line with our theoretical result in Proposition 1.4, the R_k^2s for the full sample are decreasing with increasing forecast horizon. As expected, R_1^2 is below the upper bound of one-third (see Equation (1.13)). Among the three regimes, we observe the highest R_k^2s in the high volatility regime. Clearly, the high R_k^2s in the high volatility regime do not reflect an improved absolute forecast performance

[15] In contrast, in the empirical forecast evaluation in Subsection 1.4.4 we apply a "rolling scheme." As we will discuss below, this is important because it takes into account the real-time nature of the data and allows for changes in the model parameters.

[16] In-sample parameter estimates for the MS-GARCH-TVI model can be found in the Appendix, Table 1.9. The median estimates of α and β are close to the true values. The estimates of ω_1 and ω_2 represent a low and a high volatility regime. As measured by $\varpi = p_{1,1} + p_{2,2} - 1$, the degree of persistence in the long-term component is very high.

Table 1.1: Monte-Carlo parameter estimates.

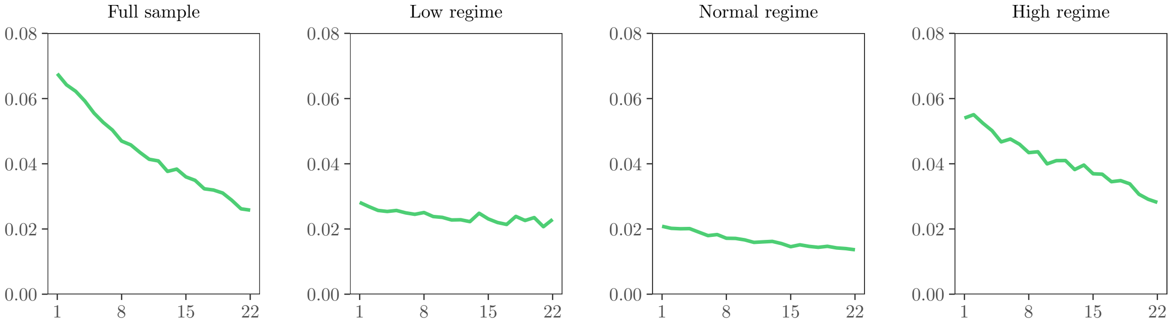
		α	β	m	θ	w_2	$\kappa-3$
		Panel A: $Z_{n,i,t}$ normally distributed					
Monthly τ_t	GARCH-MIDAS (36)	-0.000	-0.004	-0.007	0.036	1.959	-0.010
		{0.008}	{0.014}	{0.071}	{0.145}	{6.494}	
		(0.009)	(0.015)	(0.070)	(0.137)	(12.240)	
	GARCH-MIDAS (12)	-0.000	-0.003	-0.006	-0.029	-0.470	-0.009
	GARCH-MIDAS (36, $\tilde{X}$)	0.000	-0.003	-0.006	0.000	0.788	-0.009
	GARCH-MIDAS (12, $\tilde{X}$)	0.000	-0.002	-0.005	-0.075	-0.869	-0.008
	GARCH	0.000	0.003	0.009	—	—	0.001
Daily τ_t	GARCH-MIDAS (264)	-0.000	-0.003	-0.003	0.010	1.030	-0.006
		{0.008}	{0.014}	{0.063}	{0.078}	{5.020}	
		(0.008)	(0.014)	(0.062)	(0.075)	(4.786)	
	GARCH-MIDAS (66)	-0.000	-0.002	-0.001	-0.053	-3.247	-0.004
	GARCH-MIDAS (264, $\tilde{X}$)	-0.000	-0.003	-0.002	0.002	0.332	-0.005
	GARCH-MIDAS (66, $\tilde{X}$)	0.000	-0.002	0.000	-0.066	-3.414	-0.003
	GARCH	0.003	0.003	0.031	—	—	0.020
		Panel B: $Z_{n,i,t}$ student-t distributed					
Monthly τ_t	GARCH-MIDAS (36)	-0.000	-0.004	-0.008	0.040	1.491	0.108
		{0.008}	{0.014}	{0.075}	{0.152}	{5.983}	
		(0.008)	(0.015)	(0.071)	(0.141)	(11.033)	
	GARCH-MIDAS (12)	-0.000	-0.003	-0.006	-0.030	-0.589	0.109
	GARCH-MIDAS (36, $\tilde{X}$)	-0.000	-0.003	-0.006	0.003	0.715	0.110
	GARCH-MIDAS (12, $\tilde{X}$)	-0.000	-0.002	-0.004	-0.073	-0.797	0.111
	GARCH	-0.000	0.003	0.011	—	—	0.122
Daily τ_t	GARCH-MIDAS (264)	-0.000	-0.003	-0.002	0.012	1.136	0.112
		{0.008}	{0.014}	{0.065}	{0.082}	{5.896}	
		(0.008)	(0.014)	(0.063)	(0.075)	(6.039)	
	GARCH-MIDAS (66)	0.000	-0.002	0.000	-0.052	-2.730	0.114
	GARCH-MIDAS (264, $\tilde{X}$)	0.000	-0.003	-0.001	0.003	0.341	0.114
	GARCH-MIDAS (66, $\tilde{X}$)	0.000	-0.002	0.001	-0.064	-3.372	0.116
	GARCH	0.003	0.003	0.034	—	—	0.141

Notes: The table reports the average bias of parameter estimates and the corresponding standard errors across 2,000 Monte-Carlo simulations. We provide results for both daily and monthly long-term components. In curly brackets, empirical standard deviations of parameter estimates are reported. Entries in parentheses correspond to the square root of average Wang and Ghysels (2015) asymptotic variances. The parameter estimates are based on (the first) 20 years of observations (i.e. the in-sample period). In both long-term components (see Equations (2.8.1) and (2.5)), we choose $\theta = 0.3$ and $w_1 = 1$. We use $m = 0.1$ and $w_2 = 4$ in the monthly τ_t and $m = -0.1$ and $w_2 = 5$ in the daily τ_t. The long-term component is assumed to depend on $K = 36$ monthly or $K = 264$ daily observations. The covariate X_t is modeled as an AR(1) process; that is, $X_t = \phi X_{t-1} + \xi_t, \xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \sigma_\xi^2)$, with $\phi = 0.9$, $\sigma_\xi^2 = 0.3^2$ for a monthly, and $\phi = 0.98$, $\sigma_\xi^2 = 0.2^2$ for a daily τ_t. The parameters of the short-term component are in both cases given by $\alpha = 0.06$, $\beta = 0.91$ and $\gamma = 0$. For each model that is estimated based on the true value of X_t, we also incorporate estimations in which X_t is replaced by a noisy proxy $\tilde{X}_t$. It is modeled as $\tilde{X}_t = X_t + \mathcal{N}(0, 0.2 + 0.8|X_t|)$ in the case of the monthly varying τ_t and $\tilde{X}_t = X_t + \mathcal{N}(0, 0.5 + 0.8|X_t|)$ in the case of a daily varying τ_t. The column "$\kappa - 3$" presents the mean excess kurtosis of the standardized residuals from each model.

but rather an improved relative forecast performance. Further, note that for almost all forecast horizons the R^2_ks in the full sample are higher than in each subsample.

For empirical applications, cumulative volatility forecasts are of greater importance than k-step-ahead forecasts. Hence, in Figure 1.5 we present the $R^2_{1:k}$ of MZ regressions for cumulative volatility forecasts, $h_{1:k,t+1|t}$, with $k = 1, \ldots, 22$. Note that, by construction, the volatility forecasts are non-overlapping. We now present forecasts from the correctly specified and the mis-

Figure 1.4: MZ R^2—monthly τ_t—evaluation based on $\varepsilon^2_{k,t+1}$.



Notes: The figure shows the average R^2_k of MZ regressions based on the predictions from the correctly specified GARCH-MIDAS model over all 2,000 Monte Carlo replications. The true volatility is proxied by $\varepsilon^2_{k,t+1}$. Besides the full out-of-sample period, we consider low-, normal-, and high-volatility regimes. For the definition of the regimes see Subsection 1.3.3.

specified GARCH-MIDAS models as well as from the MS-GARCH-TVI and the nested GARCH. Forecast evaluation is based on the precise proxy $RV_{1:k,t+1}$. Panels (a)/(b) show the results for monthly/daily τ_t. Based on Figure 1.5, we are able to rank the different models' forecast performance. While the performance of all GARCH-MIDAS models is essentially indistinguishable, the one-component GARCH and the MS-GARCH-TVI models lead to a lower $R^2_{1:k}$. Differences between models are most pronounced in the low and normal regime.

Model confidence sets

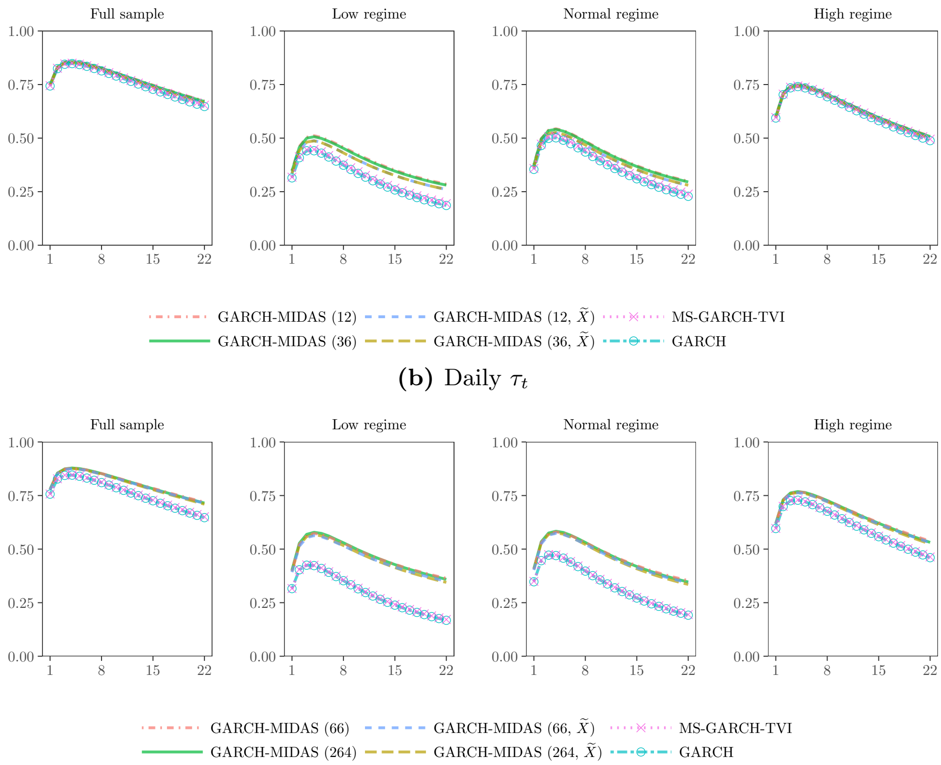
Next, we formally test for superior predictive ability. We base our analysis on the MCS approach introduced by Hansen, Lunde, and Nason (2011). Following the arguments in Patton (2011), we use the QLIKE loss as the evaluation criterion. For a k-step-ahead volatility forecast, the QLIKE is defined as

$$\text{QLIKE}\left(\sigma^2_{k,t+1}, h_{k,t+1|t}\right) = \sigma^2_{k,t+1}/h_{k,t+1|t} - \log\left(\sigma^2_{k,t+1}/h_{k,t+1|t}\right) - 1.$$

The QLIKE is the only robust loss function that depends solely on the standardized forecast error, $\sigma^2_{k,t+1}/h_{k,t+1|t}$. As discussed in Patton (2011), the QLIKE is less sensitive with respect to extreme observations than the squared error loss. Further, it can be shown that the moment conditions required for Diebold and Mariano (1995) or Giacomini and White (2006) type tests are weaker under QLIKE than under squared error loss (Patton, 2006).

We consider the following forecasting schemes. Based on the information available at the last

Figure 1.5: MZ $R^2_{1:k}$—monthly and daily τ_t—evaluation based on $RV_{1:k,t+1}$.



(a) Monthly τ_t

(b) Daily τ_t

Notes: For each model the figure shows the average $R^2_{1:k}$ of the MZ regressions over the 2,000 Monte Carlo replications. The true volatility is proxied by $RV_{1:k,t+1}$. The upper/lower panels display the case of monthly/daily long-term components. Besides the full out-of-sample period, we consider low-, normal-, and high-volatility regimes. For the definition of the regimes see Subsection 1.3.3.

day of the current month, cumulative volatility forecasts are computed for horizons of one day (1d), two weeks (2w) and one month (1m) as well as forecasts of volatility in two months (2m) and three months (3m). Whenever the forecast horizon is longer than the frequency of the long-term component, the optimal forecast requires predicting the long-term component. Instead, we simply fix the long-term component at its current level (see Subsection 1.2.4). Forecast evaluation is now based on the precise proxy $RV_{1:k,t+1}$. Next, we explain how the MCS is obtained. Denote by $\mathcal{M}$ the set of all competing models. We define

$$d_{i,j}(s,k) = \text{QLIKE}(RV_{1:k,t+s}, \hat{h}^{(i)}_{1:k,t+s|t}) - \text{QLIKE}(RV_{1:k,t+s}, \hat{h}^{(j)}_{1:k,t+s|t})$$

as the difference in the QLIKE loss of models i and j. For example, when $s = 1$ and $k \in \{1, 5, 22\}$ the forecast $\hat{h}^{(i)}_{1:k,t+s|t}$ denotes the cumulative forecast for the first (1d), the first five (1w), or all twenty-two (1m) days in the following month while for $s \in 2, 3$ and $k = 22$ we obtain the forecast for two (2m) and three (3m) months in the future. We compute the average loss difference, $\bar{d}_{i,j}$, and calculate the test statistic

$$t_{ij} = \bar{d}_{i,j} / \sqrt{\widehat{\mathbf{Var}}\left(d_{i,j}\right)} \text{ for all } i, j \in \mathcal{M}.$$

The MCS test statistic is then given by $T_{\mathcal{M}} = \max_{i,j \in \mathcal{M}} |t_{i,j}|$ and has the null hypothesis that all models have the same expected loss. Under the alternative, there is some model i that has an expected loss greater than the expected loss of all other models $j \in \mathcal{M} \setminus i$. If the null hypothesis is rejected, the worst performing model is eliminated. The test is performed iteratively, until no further model can be eliminated. We denote the final set of surviving models by $\mathcal{M}_{MCS}$. This final set contains the best forecasting model with confidence level $1 - \nu$. We set $\nu = 0.1$. This choice is common practice in the literature (e.g., Laurent, Rombouts, and Violante, 2013; Liu, Patton, and Sheppard, 2015).

Since the asymptotic distribution of the test statistic $T_{\mathcal{M}}$ is nonstandard, we approximate it by block-bootstrapping as proposed by Hansen, Lunde, and Nason (2011), where the block length is determined by fitting an AR(p) process to the series of loss differences. In our analysis, 8,000 bootstrap replications at each stage were sufficient in order to obtain stable results.[17]

Table 1.2 reports how often a certain model is included in the MCS across the 2,000 replications. Panel A provides results for normally distributed innovations and Panel B for Student's t distributed innovations. For example, for normally distributed innovations, monthly τ_t, and a forecast horizon of one day, the correctly specified GARCH-MIDAS (36) is included in the MCS in 85% of the replications. The table clearly shows that the misspecified one-component GARCH model is included less often in the MCS than the GARCH-MIDAS models. In particular, this is the case for daily τ_t. Further, for daily τ_t and forecast horizons of up to two months the MS-GARCH-TVI is less often part of the MCS than all GARCH-MIDAS models. Additionally, among the GARCH-MIDAS models the correctly specified one has the highest inclusion rates in the MCS when the forecast horizon is up to one month. At least for monthly τ_t, it appears that a misspecification of the lag length is less severe than observing the explanatory variable with measurement error. Finally, at the longest forecast horizon (3m) all forecasts suffer from a misspecified forecast of the long-term component and hence it becomes increasingly difficult to

[17] For implementing the MCS procedure, we use the R package *rugarch* (Ghalanos, 2018) which includes the implementation used in the MFE Matlab Toolbox by Kevin Sheppard. See: https://www.kevinsheppard.com/MFE_Toolbox.

distinguish between models.

Table 1.2: Model confidence set inclusion rates.

		1d	2w	1m	2m	3m
	Panel A: $Z_{n,i,t}$ normally distributed					
Monthly τ_t	GARCH-MIDAS (36)	0.850	0.758	0.770	0.795	0.792
	GARCH-MIDAS (12)	0.852	0.745	0.762	0.818	0.827
	GARCH-MIDAS (36, $\tilde{X}$)	0.723	0.559	0.589	0.650	0.661
	GARCH-MIDAS (12, $\tilde{X}$)	0.696	0.539	0.560	0.648	0.684
	MS-GARCH-TVI	0.765	0.560	0.603	0.664	0.673
	GARCH	0.477	0.221	0.216	0.260	0.310
Daily τ_t	GARCH-MIDAS (264)	0.946	0.893	0.861	0.784	0.743
	GARCH-MIDAS (66)	0.850	0.796	0.836	0.890	0.878
	GARCH-MIDAS (264, $\tilde{X}$)	0.843	0.672	0.646	0.663	0.688
	GARCH-MIDAS (66, $\tilde{X}$)	0.763	0.614	0.664	0.778	0.831
	MS-GARCH-TVI	0.376	0.100	0.138	0.467	0.765
	GARCH	0.257	0.043	0.050	0.244	0.493
	Panel B: $Z_{n,i,t}$ student-t distributed					
Monthly τ_t	GARCH-MIDAS (36)	0.912	0.790	0.772	0.761	0.764
	GARCH-MIDAS (12)	0.922	0.808	0.785	0.812	0.818
	GARCH-MIDAS (36, $\tilde{X}$)	0.842	0.656	0.640	0.652	0.650
	GARCH-MIDAS (12, $\tilde{X}$)	0.841	0.636	0.622	0.668	0.683
	MS-GARCH-TVI	0.875	0.666	0.654	0.675	0.664
	GARCH	0.734	0.331	0.267	0.280	0.309
Daily τ_t	GARCH-MIDAS (264)	0.968	0.912	0.866	0.792	0.742
	GARCH-MIDAS (66)	0.918	0.839	0.862	0.885	0.854
	GARCH-MIDAS (264, $\tilde{X}$)	0.927	0.769	0.712	0.694	0.685
	GARCH-MIDAS (66, $\tilde{X}$)	0.877	0.726	0.731	0.812	0.822
	MS-GARCH-TVI	0.690	0.222	0.206	0.501	0.758
	GARCH	0.602	0.112	0.093	0.276	0.485

Notes: The numbers are the empirical frequencies of a model being included in the 90% model confidence set at different forecast horizons: one day (1d), two weeks (2w), one month (1m), two months (2m), and three months (3m). Panel A corresponds to the simulation with normally distributed intraday returns and Panel B to standardized Student's t distributed intraday returns with five degrees of freedom. The averages are taken across 2,000 Monte-Carlo replications.

In summary, independently of whether the long-term component is specified at a daily or monthly frequency, the correctly specified GARCH-MIDAS model as well as the GARCH-MIDAS with misspecified lag length clearly outperform the one-component GARCH as well as the MS-GARCH-TVI in terms of forecast performance. For models with daily long-term components this result also holds when the explanatory variable is observed with measurement error. Only for monthly long-term components and measurement error in X_t, we find that the MS-GARCH-TVI performs slightly better.

Remark 1.2. *As discussed in Subsection 1.2.1, Assumption 1.3 is likely to hold for explanatory variables that are observed at a lower frequency than the daily returns. For certain daily* *explanatory variables (e.g. the VIX index) Assumption 1.3 might be violated. However, under reasonable assumptions the correlation between the innovations to the daily returns and X_t itself can be expected to be small. The correlation with future τ_t will be even smaller. For a more detailed discussion see Appendix 1.6.4, which also provides additional simulations. The simulations show that even if Assumption 1.3 is mildly violated all the previous findings still hold.*

1.4 Empirical analysis

Last, we turn to an empirical application of the GARCH-MIDAS models to S&P 500 return data. In Subsection 1.4.1 we introduce our data set. Full sample estimation results for various GARCH-MIDAS models are reported in Subsection 1.4.2. Thereafter, in Subsection 1.4.3 we explain how real-time volatility forecasts can be constructed when taking into account the release schedule of macroeconomic variables. The forecast comparison is carried out in Subsection 1.4.4, where we evaluate the GARCH-MIDAS volatility forecasts against forecasts from eight competitor models.

1.4.1 Data

Stock market data

We consider daily log-returns on the S&P 500, calculated as $r_{i,t} = 100 \cdot (\log(p_{i,t}) - \log(p_{i-1,t}))$, for the 1971:M1 to 2018:M4 period. For evaluating the volatility forecasts, we employ daily realized variances, $RV_{i,t}$, defined as the sum of the squared five-minute intraday log-returns on day t plus the squared overnight log-return. The latter is defined as the log of the open price on day t minus the log of the close price on day $t-1$. This approach follows Bollerslev et al. (2018), among others. The data for constructing $RV_{i,t}$ were obtained from the Realized Library of the Oxford-Man Institute of Quantitative Finance and are available from the year 2000 onwards (Heber et al., 2009).

Explanatory variables

As explanatory variables we use daily measures of financial risk, a weekly measure of financial conditions and monthly macroeconomic variables. We employ backward- and forward-looking measures of daily volatility. The former is proxied by a rolling window of the average realized volatility (based on squared daily returns) over the previous 22 days, $RVol(22)_{i,t} = \sqrt{1/22 \sum_{j=0}^{21} r_{i-j,t}^2}$, and the latter by the VIX index (converted to a daily level by dividing it by $\sqrt{252}$). In addition, we consider the difference between the VIX (divided by $\sqrt{252}$) and RVol(22) as a proxy for the (square root of the) variance risk premium (VRP).[18]

We use the weekly National Financial Conditions Index (NFCI) as a measure for the tightness of financial conditions in the USA. The NFCI is a weighted average of 105 standardized financial indicators of risk, credit and leverage derived by dynamic factor analysis. Monthly macroeconomic conditions are measured by the Chicago Fed National Activity Index (NAI) and growth rates of industrial production and housing starts, both calculated as $\Delta X_t = 100 \cdot (\log(X_t) - \log(X_{t-1}))$. While the macroeconomic variables are included from 1971 onwards, the NCFI series begins in 1973 and the VIX is available from 1990 onwards.[19]

Before we estimate GARCH-MIDAS models, we employ the Conrad and Schienle (2020) Lagrange multiplier (LM) test for an omitted multiplicative component in one-component GARCH models. This test checks whether a simple GJR-GARCH(1,1) is misspecified in the sense of neglecting a second component that is driven by an explanatory variable X. Since the test is of the LM type, it requires estimation of the model under the null hypothesis only. Assuming that under the alternative there is a second component which is driven by K lags of the variable X, the test statistic can be shown to be χ^2 with K degrees of freedom. An appealing property of the test is that it can be applied in settings where X is observed at the same frequency as the returns but also when X is observed at a lower frequency. Intuitively, the test checks whether the squared standardized residuals from the GJR-GARCH are predictable using (functions of) past values of X. Table 1.3 shows the outcome of the test when applied to each of our explanatory variables. When either choosing $K = 1$ or $K = 2$, the test clearly rejects the null hypothesis that a GJR-GARCH is correctly specified for all variables except housing starts. Thus, the LM test results suggest using GARCH-MIDAS models instead. The estimates for a GARCH-MIDAS model based on housing starts in Subsection 1.4.2 will show that housing starts are a leading indicator with respect to financial volatility. This implies that the choice of $K = 1$ or $K = 2$ is too small. When redoing the LM test for a lag length of up to $K = 12$ the LM test indeed

[18] Note that the conventional definition of the variance risk premium is the squared VIX minus realized variance. We are interested in expressing the quantity in volatility units. Because the realized VRP takes positive as well as negative values, we take the square root of both quantities before we take the difference.

[19] Table 1.10 in the Appendix provides summary statistics for the stock returns and the seven explanatory variables. Figure 1.8 in the Appendix shows the evolution of the corresponding time series. Further details on the data set are provided in Appendix 1.6.6.

rejects the null hypothesis also for housing starts.

Table 1.3: LM test for misspecification of GJR-GARCH(1,1).

X_t	VIX	RVol(22)	NFCI	NAI	Δ IP	Δ Housing
$K = 1$	76.28	14.38	22.54	15.25	7.99	0.18
	[<0.01]	[<0.01]	[<0.01]	[<0.01]	[<0.01]	[0.67]
$K = 2$	84.05	19.03	24.05	17.34	10.22	0.18
	[<0.01]	[<0.01]	[<0.01]	[<0.01]	[<0.01]	[0.91]

Notes: The table reports the test statistics and the corresponding p-values of the Conrad and Schienle (2018) misspecification test for one-component GJR-GARCH(1,1) models. The test is implemented using either one ($K = 1$) or two ($K = 2$) lags of the explanatory variable X_t. For VIX and RVol(22) the test is based on daily data from 1990 onwards, for NFCI, NAI, Δ IP, and Δ Housing starts the test is based on weekly/monthly data from 1974 onwards.

We can also apply the LM test jointly to several variables at the same time. However, all variables need to be observed at the same frequency. When including the NAI, industrial production and housing starts and selecting an appropriate lag length, the NAI and housing starts are individually significant while industrial production is not. This suggests that among the macroeconomic variables the NAI and housing starts are most informative. We also aggregated the VIX and the NFCI to a monthly frequency and performed the LM test jointly for all variables. While the overall LM statistic is highly significant, the VIX, the NFCI and housing starts are the only variables that are individually significant.

1.4.2 Full sample parameter estimates

One explanatory variable

We first estimate a GARCH-MIDAS model for each explanatory variable for the full sample. We include a constant in the mean equation; that is, returns are modeled as $r_{i,t} = \mu + \varepsilon_{i,t}$. After visual inspection of the estimated weighting schemes for alternative choices of K, we select a lag length that is rather too large than too small. As discussed in Section 1.3, the data will identify the optimal weighting scheme as long as K is chosen large enough. We choose $K = 264$ for RVol(22), $K = 3$ for the VIX/VRP and $K = 52$ for the NFCI.[20] Thus, for the forward-looking VIX/VRP only the most recent information appears to drive long-term volatility, while the backward-looking RVol(22) is smoothed over many lags. As in Conrad and Loch (2015), we choose $K = 36$ for the monthly macroeconomic variables. The estimates for the parameters

[20] For all variables, Figure 1.9 in the Appendix shows the estimated weighting schemes for selected choices of K. The figure illustrates that the estimated weighting schemes no longer change once the selected lag length is sufficiently large. In all cases, our choice of the lag length is rather conservative.

in the conditional variance are reported in Table 1.4. For all variables except housing starts, we find that a restricted Beta weighting scheme with $w_1 = 1$ is the best choice; that is, the optimal weights are declining from the beginning. For housing starts, an unrestricted scheme which allows for "hump-shaped" weights is required. This confirms the finding in Conrad and Loch (2015) that housing starts are leading with respect to long-term volatility.[21] Note that the GARCH-MIDAS models based on the NFCI and the three macroeconomic variables employ return data for the 1974:M1 to 2018:M4 period, while the models with daily τ_t employ data from 1990:M1 onwards. Hence models based on daily τ_t cannot be compared to models based on weekly/monthly τ_t in terms of log-likelihood or Bayesian Information Criterion (BIC).

Table 1.4: Full sample estimation results: GARCH-MIDAS with one explanatory variable.

	α	β	γ	m	θ	w_1	w_2	K	LLH	BIC	VR(X)
Daily τ_t											
RVol(22)	0.000 (0.008)	0.843*** (0.012)	0.192*** (0.015)	−1.261*** (0.112)	1.177*** (0.096)	1	3.049*** (0.675)	264	−9201	18465	42.78
VIX	0.000 (0.010)	0.853*** (0.021)	0.095*** (0.015)	−2.129*** (0.086)	1.524*** (0.067)	1	3.470** (1.371)	3	−9138	18339	76.14
VRP	0.017** (0.007)	0.902*** (0.007)	0.128*** (0.011)	−0.384*** (0.137)	1.084*** (0.096)	1	5.571** (2.591)	3	−9174	18410	10.92
Weekly τ_t											
NFCI	0.017*** (0.006)	0.902*** (0.005)	0.115*** (0.007)	−0.101 (0.073)	0.252*** (0.048)	1	2.892 (2.314)	52	−15103	30271	11.42
Monthly τ_t											
NAI	0.019*** (0.006)	0.900*** (0.005)	0.116*** (0.007)	−0.058 (0.079)	−0.359*** (0.073)	1	9.066*** (3.312)	36	−14569	29202	14.14
Δ IP	0.019*** (0.006)	0.903*** (0.005)	0.113*** (0.007)	0.074 (0.089)	−0.650*** (0.161)	1	5.271*** (1.782)	36	−14573	29211	10.63
Δ Housing	0.019*** (0.005)	0.897*** (0.005)	0.119*** (0.007)	−0.079 (0.076)	−0.237*** (0.034)	1.695*** (0.383)	2.586*** (0.770)	36	−14559	29192	19.63
GARCH	0.021*** (0.005)	0.911*** (0.005)	0.103*** (0.007)	−0.073 (0.098)	—	—	—	—	−15355	30757	—

Notes: Estimation results for GARCH-MIDAS models are reported for seven explanatory variables. The estimation using the NFCI, NAI, IP, and housing starts begins in 1974:M1 based on low-frequency observations reaching as far as 1971:M1 in line with the lag length K. The estimation of the GARCH-MIDAS models using RVol(22) and VIX as an explanatory variable employs daily return data starting in 1990:M1. For all explanatory variables except housing starts a restricted weighting scheme is chosen ($w_1 = 1$). Bollerslev-Wooldridge standard errors are reported in parentheses where significance at the 1, 5, 10 % level is indicated by ***, **, and *. LLF is the value of the maximized log-likelihood function and BIC is the Bayesian Information Criterion. The variance ratio VR(X) = $\mathbf{Var}(\log(\tau_M^X))/\mathbf{Var}(\log(\sigma_M^X))$ is calculated on monthly aggregates. Estimates for μ are omitted.

Concerning the parameter estimates, it is interesting to observe that the GARCH-MIDAS models with daily τ_t lead to lower estimates of β than models with weekly or monthly τ_t. While for the models with daily τ_t the estimates of α are close to zero, there is strong evidence for asymmetry (as indicated by the highly significant γ parameter). These parameter estimates imply that the deviations of the short-term component from the long-term component are more

[21] Figure 1.10 in the Appendix shows the estimated weighting schemes.

short-lived for GARCH-MIDAS models with daily τ_t.[22] The signs of the estimated θs for realized volatility, the VIX, and the macroeconomic variables are in line with findings in the previous literature. Higher levels of financial volatility tend to increase long-term volatility, whereas an improvement in macroeconomic conditions decreases long-term volatility. The finding that a higher variance risk premium and tighter financial conditions (i.e., an increase in the NFCI) predict higher volatility is new. While the positive relation between realized/expected measures of volatility and long-term volatility might be viewed as "mechanical," the NFCI as well as the macroeconomic variables can be considered fundamental drivers of financial volatility.

We gauge the importance of the variation in the long-term component for the overall expected variation in return volatility by the variance ratio introduced in Equation (1.7). To facilitate comparison across models, we focus on the monthly variation of volatility. That is, for all models we denote the monthly aggregate volatility by σ_M^X. For models with monthly long-term components, we have that $\tau_M^X = \tau_t^X$. For models with daily or weekly long-term components, τ_M^X refers to monthly aggregates of the daily/weekly long-term component. We then calculate $\mathrm{VR}(X) = \mathbf{Var}(\log(\tau_M^X))/\mathbf{Var}(\log(\sigma_M^X))$, where X indicates that the variance ratio is based on a specific explanatory variable. As Table 1.4 shows, the models with daily τ_t achieve much higher variance ratios than the models with a weekly/monthly long-term component. Among the models with daily long-term components, the variance ratio of 76.14% for the VIX-based model is by far the highest and implies that three quarters of the expected variation in return volatility can be traced back to variation in the VIX. In Section 1.4.4 we will investigate whether a high variance ratio necessarily translates into good OOS predictive performance.

Two explanatory variables

The GARCH-MIDAS setting allows us to include two or more explanatory variables in the long-term component. Based on the results in the previous section, the VIX appears to be better suited to capture daily movements in the long-term component than RVol(22) or the VRP. Since the NFCI and, in particular, the macroeconomic variables capture lower frequency movements, it is natural to estimate GARCH-MIDAS models with the VIX and one of those variables jointly in the long-term component. This allows us to formally check whether the NFCI and the three macroeconomic variables contain information that is complementary to the VIX. The long-term

[22] This behavior is also evident from Figure 1.11 in the Appendix which shows the evolution of the annualized long-term components and the conditional volatilities.

component for those models is given by:

$$\log \tau_{i,t} = m + \theta^{VIX} \sum_{l=1}^{K^{VIX}} \varphi_l(1, w_2^{VIX}) VIX_{i-l,t} + \theta^{X} \sum_{l=1}^{K^{X}} \varphi_l(w_1^{X}, w_2^{X}) X_{t-l}.$$

Estimation results are presented in Table 1.5. Note that K^{VIX} and K^X are chosen as in Table 1.4. For all models the estimation period is now determined by the availability of the VIX. When controlling for the VIX, the θ^X parameter turns out to be significant for the NAI and housing starts. Thus, macroeconomic variables appear to contain information that is complementary to the one included in the VIX. However, none of the models that include two variables achieves a higher VR than the model based on the VIX alone.

Table 1.5: Full sample estimation results: VIX combined with second explanatory variable.

	α	β	γ	m	θ^X	w_1^X	w_2^X	θ^{VIX}	w_2^{VIX}	K^X	LLH	BIC	VR(VIX, X)
Daily τ_t													
VIX	0.000 (0.010)	0.853*** (0.021)	0.095*** (0.015)	−2.129*** (0.086)	—	—	—	1.524*** (0.067)	3.470** (1.371)	3	−9138	18339	76.14
Weekly τ_t													
NFCI	0.000 (0.010)	0.852*** (0.020)	0.099*** (0.016)	−1.993*** (0.143)	0.118 (0.085)	1	2.252 (4.152)	1.451*** (0.093)	3.617** (1.518)	52	−9110	18300	75.84
Monthly τ_t													
NAI	0.000 (0.009)	0.870*** (0.018)	0.092*** (0.015)	−2.032*** (0.100)	−0.108** (0.046)	1	119.372 (326.330)	1.431*** (0.079)	3.775** (1.594)	36	−9133	18346	75.06
Δ IP	0.000 (0.009)	0.876*** (0.018)	0.084*** (0.014)	−2.133*** (0.096)	−0.043 (0.089)	1	8.960 (34.803)	1.528*** (0.072)	3.806** (1.520)	36	−9139	18357	75.91
Δ Housing	0.000 (0.009)	0.863*** (0.019)	0.097*** (0.015)	−2.035*** (0.094)	−0.061** (0.024)	1.001 (0.743)	2.139 (2.462)	1.446*** (0.074)	3.605** (1.503)	36	−9135	18359	74.99

Notes: Estimation results for GARCH-MIDAS models are reported in which the daily VIX is combined with the low-frequency variables reported in Table 1.4—that is, the NFCI, NAI, and changes in industrial production and housing starts. The estimates are based on daily return data from 1990:M1 to 2018:M4. For comparison, the estimation results using only the VIX as a covariate from Table 1.4 are included in the first row. All parameters with a superscript X relate to the second explanatory variable. K^{VIX} is always equal to three. Bollerslev-Wooldridge standard errors are reported in parentheses where significance at the 1, 5, 10 % level is indicated by ***, **, and *. LLF is the value of the maximized log-likelihood function and BIC is the Bayesian Information Criterion. The variance ratio $\text{VR}(VIX, X) = \mathbf{Var}(\log(\tau_M^{VIX,X}))/\mathbf{Var}(\log(\sigma_M^{VIX,X}))$ is calculated on monthly aggregates. Estimates for μ are omitted.

More than two explanatory variables

As an extension to Subsection 1.4.2, one could employ more than two covariates. We experimented with combining three variables in the long-term component but found no further improvements in terms of model fit. Moreover, GARCH-MIDAS models including more than two variables in the long-term component are difficult to estimate because the likelihood is relatively insensitive with respect to changes in the weighting parameters. Instead, in Subsection 1.4.4 on OOS forecasting, we will aggregate the information in the different variables by simply calculating the average forecast across all GARCH-MIDAS models with one explanatory variable.

1.4.3 Real-time estimates

In the following, we make use of vintage data. This allows for a realistic evaluation of the GARCH-MIDAS models' ability to describe the behavior of long-term financial volatility in real time.[23] In order to compare full-sample estimates of the long-term component with corresponding real-time estimates, we reestimate all GARCH-MIDAS models from Table 1.4 on a daily basis. Estimation is performed on a rolling window. For each explanatory variable, the window size is determined by the length of the first estimation period ending in 2009:M12. The period 2010:M1 to 2018:M4 will be used as the OOS period for the forecast evaluation in Section 1.4.4. In order to ensure that our estimates of the long-term component are feasible in real time, we employ vintage data that is available for the NFCI, the NAI, IP, and housing starts from the ALFRED database hosted by the St. Louis Fed.[24] When using real-time data, the long-term component no longer changes its value at the beginning of a week/month but whenever a new data release becomes available.

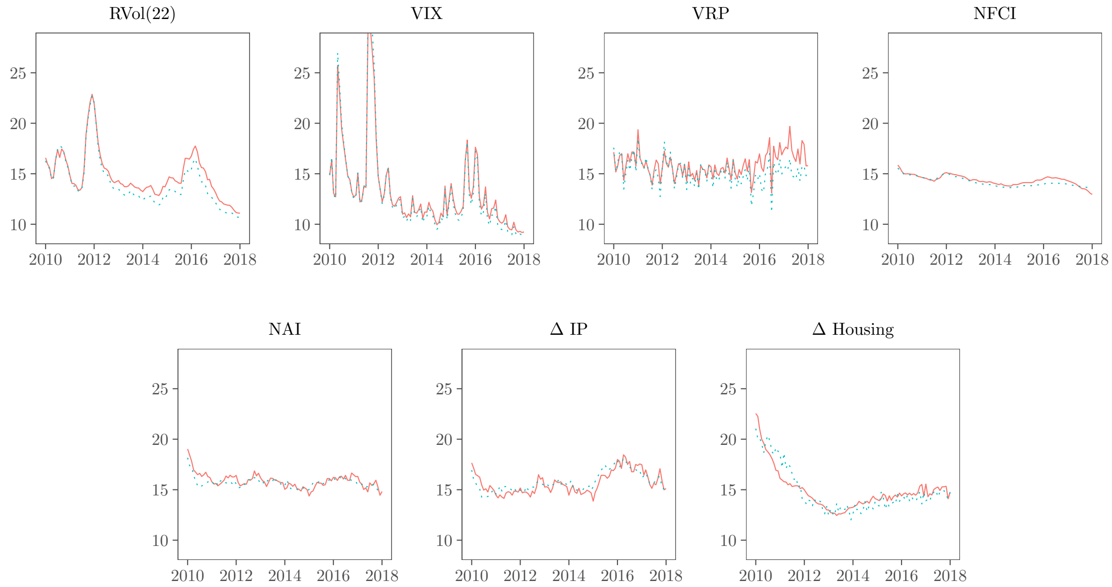
Figure 1.6 shows the estimated long-term components based on the full sample estimates (as reported in Table 1.4, dotted lines) and based on the rolling window real-time estimates (solid lines). For RVol(22), the VIX, and the VRP the long-term component estimates in the full sample might differ from the rolling window estimates, because they are based on distinct sample periods (rolling window vs. full sample). For the NFCI, the NAI, IP, and housing starts, the two long-term components are not only based on distinct sample periods but also on different data vintages (real-time vs. final). Figure 1.6 shows that for RVol(22), the VIX, and the VRP the rolling window estimate of the long-term component is often somewhat higher than the full-sample estimate. For the macroeconomic variables the real-time estimates of the long-term component are occasionally below or above the full-sample estimates. However, the average absolute differences are quite sizable. For example, the average absolute difference between the full-sample and real-time estimates based on industrial production is 6.80%. To put this into context, for industrial production the mean absolute revision from the initial release to the latest available data was 2.18% during the 1965:Q3 to 2006:Q4 period (Croushore, 2011). Among the variables considered in Croushore (2011), this is the highest value (even higher than for GDP).

[23] To the best of our knowledge, Lindblad (2017) appears to be the only other paper that makes use of real-time data when estimating GARCH-MIDAS models.

[24] For more details on real-time data availability see Appendix 1.4.3.

Similar numbers in terms of changes in the long-term component are obtained for the other variables: 9.35% for housing starts, 4.78% for the NAI, and 2.68% for the NFCI. In summary, these figures highlight the importance of using real-time instead of final data releases for the macroeconomic variables for a realistic forecast evaluation.

Figure 1.6: Comparison of rolling window and full sample long-term components.



Notes: For each explanatory variable, the monthly averaged long-term volatility components, $\sqrt{\tau_t}$, are depicted for the period 2010:M1 to the end of 2018:M1, the last month of issuing forecasts and, hence, real-time estimation. The long-term component obtained from the full sample estimates is given in green (dotted). Real-time estimates of the most recently fitted $\sqrt{\tau_t}$ are depicted in red (solid). Volatilities are presented on an annualized scale.

1.4.4 Forecast evaluation

Finally, we evaluate the predictive performance of the GARCH-MIDAS models in the 2010:M1 to 2018:M4 OOS period. As before, we consider cumulative volatility forecasts for horizons up to three months. When computing the forecasts, we keep the long-term component fixed at its current level. Volatility forecasts are based on the real-time rolling window parameter estimates as obtained in Subsection 1.4.3 (i.e., we apply a “rolling (forecasting) scheme”).

Competitor models

For forecast comparison, we use an extensive range of competitor models which are either extensions of the simple GARCH specification or which model the realized variance directly.

First, we consider the simple one-component GARCH(1,1) model and a no-change (or randomwalk) forecast which simply scales the realized variance on the last day of period t to the appropriate horizon: $h_{1:k,t+s|t} = k \cdot RV_{n,t}$. Second, we use the MS-GARCH-TVI model that we employed in Subsection 1.3.3. The only difference is that we now use a GJR-GARCH specification in both regimes. In addition, we use an MS-GARCH model that consists of two GARCH equations with individual intercepts and individual ARCH and GARCH parameters. We incorporate asymmetric effects in the low volatility regime only.[25] We refer to this model as MS-GARCH with time-varying coefficients (MS-GARCH-TVC). Further, we use the HEAVY model by Shephard and Sheppard (2010) and the Realized GARCH model by Hansen, Huang, and Shek (2012). The specifications of the HEAVY and the Realized GARCH models employ a measure of pure intraday realized variance, $RV_{i,t}^{int}$ (defined as the sum of squared intraday returns). Third, we consider two specifications that directly model the realized variance, $RV_{i,t}$, (including squared overnight returns) and allow us to compute direct (as compared to iterated) volatility forecasts. We employ the HAR model of Corsi (2009) and the HAR model with leverage effect proposed in Corsi and Renò (2012).

For more details on the exact specification of the competitor models, their estimation and volatility forecasting see Appendix 1.6.7.[26] For the OOS forecast evaluation all competitor models are reestimated on a rolling window basis.

Forecast error statistics and model confidence set

As in Subsection 1.3.3, we base the comparison of the forecast performance of the different models on the QLIKE loss. Table 1.6 reports the average QLIKE loss for each model and forecast horizons of one day (1d), two weeks (2w), one month (1m), two months (2m), and three months (3m). We use the MCS approach to test whether there is one or several models that significantly outperform the others. As in Subsection 1.3.3, we rely on 90% model confidence sets.[27]

MCS for full OOS period. Blue areas in Table 1.6 indicate that for the corresponding forecast horizon the respective model is included in the final set, $\mathcal{M}_{MCS}$. For example, for a forecast horizon of one day the only model that is included in the final MCS is the HAR model with

[25] Initially, we estimated a GJR-GARCH specification in both regimes. However, it turned out that the asymmetry term was only significant in the component which represents the low volatility regime. In addition, we select this specification because it is much more stable in the rolling window estimation than the one with two GJR-GARCH regimes.

[26] Table 1.11 in the Appendix shows the full sample parameter estimates for the competitor models.

[27] As a robustness check, we present the corresponding results for a 95% MCS in Appendix 1.6.8. Essentially all findings remain unaffected.

Table 1.6: QLIKE losses and model confidence sets: full out-of-sample period.

	1d	2w	1m	2m	3m
GARCH-MIDAS					
RVol(22)	0.306	0.246	0.271	0.387	0.428
VIX	0.275	0.215	0.240	0.359	0.414
VRP	0.291	0.227	0.260	0.384	0.430
NFCI	0.324	0.248	0.264	0.363	0.393
NAI	0.343	0.266	0.283	0.391	0.424
Δ IP	0.345	0.267	0.285	0.395	0.438
Δ Housing	0.328	0.252	0.264	**0.347**	**0.380**
VIX and NFCI	0.274	0.213	0.236	0.349	0.399
VIX and NAI	0.275	0.215	0.241	0.358	0.409
VIX and Δ IP	0.274	0.214	0.239	0.355	0.409
VIX and Δ Housing	0.275	0.218	0.243	0.351	0.405
Avg.	0.317	0.246	0.264	0.364	0.400
Competitor models					
GARCH	0.342	0.263	0.282	0.395	0.434
MS-GARCH-TVI	0.362	0.292	0.315	0.426	0.488
MS-GARCH-TVC	0.355	0.271	0.283	0.387	0.421
RealGARCH	0.260	**0.206**	**0.233**	0.356	0.390
HEAVY	0.277	0.238	0.299	0.539	0.662
HAR	0.254	0.210	0.243	0.368	0.419
HAR (lev.)	**0.238**	0.207	0.245	0.371	0.419
No-change	0.358	0.498	0.636	1.157	1.292

Notes: Numbers reported are the average out-of-sample QLIKE losses for each model for one-day- (1d), two-week- (2w), one-month- (1m), two-month- (2m) and three-month-ahead (3m) variance forecasts. Bold entries indicate the model with the lowest average QLIKE loss per horizon. Blue-shaded numbers indicate that the respective model is included in the 90% model confidence set. The average forecast (avg.) is the mean forecast across all GARCH-MIDAS models employing one explanatory variable. The out-of-sample evaluation period spreads 2010:M1 to 2018:M4.

leverage. Thus, at the very short horizon of one day the HAR with leverage dominates all other models. At forecast horizons of two weeks the MCS includes both HAR models, the Realized GARCH, and GARCH-MIDAS specifications that either include the VIX alone or in combination with the NFCI/NAI/IP. At the one-month horizon only the Realized GARCH and the GARCH-MIDAS that combines the VIX and the NFCI are included. The picture changes at horizons of two and three months. At these horizons GARCH-MIDAS models that either combine the VIX with the NFCI/housing starts or models based on housing starts alone are included in the MCS. These results illustrate that the performance of a GARCH-MIDAS model strongly depends on choosing the best horizon-specific explanatory variable. In summary, the HAR model with leverage and the Realized GARCH achieve the lowest QLIKE at forecast horizons of one day and two weeks/one month, respectively. In contrast, the GARCH-MIDAS model based on housing starts performs best at horizons of two and three months ahead (see the bold entries).

MCS for volatility regimes. In addition to the results for the full OOS period, we also provide MCS for subsamples of low, normal, and high volatility. We define these regimes in the same way as outlined in Subsection 1.3.3. Quantiles are now computed based on the empirical distribution of full-sample realized variances. In total, we have 764 observations in the low, 961 in the normal, and 304 in the high regime. Table 1.7 presents the regime-specific analysis.

Interestingly, in the low-volatility regime the Realized GARCH and the two HAR models are the only models in the MCS for short horizons of one day and two weeks. For a forecast horizon of one month, various GARCH-MIDAS models are included in the MCS. For three months ahead, two GARCH-MIDAS specifications based on the VIX are the only models in the MCS. The results for the normal-volatility regime are even more in favor of the GARCH-MIDAS models. At essentially all horizons GARCH-MIDAS models based on the VIX are included in the MCS. As for the full OOS period, the GARCH-MIDAS based on housing starts is the only model in the three-month MCS. Finally, in the high volatility regime and for horizons of two weeks and one month, essentially all models are included in the MCS. This result may be driven by the fact that the intermediate-term forecast performance of all models substantially deteriorates during the high-volatility regime and, therefore, it becomes increasingly difficult to distinguish between models. Nevertheless, even in the high-volatility regime the GARCH-MIDAS models are very competitive for longer forecast horizons. Specifically, GARCH-MIDAS models based on the NFCI and housing starts are included in the MCS.

In summary, we find that the informative content of the explanatory variables depends on the volatility regime. While in low- and normal- volatility regimes GARCH-MIDAS models based on the VIX or the VIX combined with another variable perform well, in high-volatility regimes models purely based on macroeconomic variables are very competitive. Because recessions typically coincide with regimes of high volatility, our results are consistent with the finding from the previous literature that macroeconomic variables are particularly useful to predict financial volatility during the onset of recessions (e.g., Paye, 2012). At the longest forecast horizons, housing starts and the NFCI become more and more important. Among the competitor models it is again the Realized GARCH which performs very well across volatility regimes.

Mincer-Zarnowitz Regressions

Lastly, we consider the outcome of MZ regressions. As Table 1.8 shows, for forecast horizons of one day and two weeks the highest R^2 is achieved by GARCH-MIDAS type models. This is in sharp contrast to the results from the previous section. However, for longer forecast horizons (1m–3m) the winning models according to the R^2 are exactly the same as when using the MCS approach. Thus, at forecast horizons at which the correct modeling of the long-term component pays off, the R^2 selects the same model as the MCS. Again, the last three columns of Table 1.8 show that the highest R^2s are obtained in the high-volatility regime.[28]

1.5 Conclusion

We introduce and discuss the properties of a class of multiplicative volatility models. This class of models includes the GARCH-MIDAS but also a variant of the MS-GARCH. We show that multiplicative volatility models can generate an autocorrelation structure in the conditional variance that mimics the long-memory-type behavior that is often observed for realized variances. We also argue that the R^2 of a MZ regression can be a misleading measure of forecast accuracy across volatility regimes because the R^2 will be the highest in the regime with the highest squared error loss. In a Monte-Carlo simulation, we investigate the properties of the QMLE of the GARCH-MIDAS model and show that the estimator is unbiased and that the Wang and Ghysels (2015) asymptotic standard errors are valid in the presence of exogenous explanatory variables. We also reveal that forecast performance is relatively insensitive with respect to

[28] For brevity, we now focus on a forecast horizon of one month.

moderate misspecification of the explanatory variable and the true lag length.

In an empirical application to S&P 500 stock returns, we compare the forecast performance of the GARCH-MIDAS model with a wide range of competitor models. As expected, relative forecast performance depends on the forecast horizon. Among all models, the HAR with leverage performs best at a one-day horizon. For longer forecast horizons the Realized GARCH is very competitive and performs best at forecast horizons of two weeks and one month. The performance of GARCH-MIDAS models depends on the choice of the explanatory variable. The best GARCH-MIDAS specifications generate volatility forecasts that are comparable to or improve upon the forecasts from the Realized GARCH. Specifically, GARCH-MIDAS specifications that combine the VIX with the NFCI are included in the MCS for forecast horizons of two weeks up to two months. Most importantly, the GARCH-MIDAS based on housing starts achieves the lowest QLIKE at forecast horizons of two and three months ahead. Thus, our results are useful for selecting the appropriate horizon-specific explanatory variable and suggest that models based on low-frequency information can be more useful than models that exploit high-frequency intraday data.

Table 1.7: QLIKE losses and model confidence sets: low/normal/high-volatility regimes.

	Low-volatility regime					Normal-volatility regime					High-volatility regime				
	1d	2w	1m	2m	3m	1d	2w	1m	2m	3m	1d	2w	1m	2m	3m
GARCH-MIDAS															
RVol(22)	0.364	0.264	0.305	0.399	0.409	0.271	0.232	0.260	0.400	0.463	0.273	0.241	0.217	0.313	0.365
VIX	0.332	0.210	0.250	0.367	0.354	0.233	0.204	0.231	0.355	0.454	0.262	0.259	0.243	0.347	0.437
VRP	0.349	0.237	0.288	0.405	0.424	0.252	0.215	0.245	0.375	0.440	0.266	0.238	0.237	0.360	0.414
NFCI	0.400	0.274	0.304	0.389	0.402	0.272	0.228	0.248	0.364	0.417	0.292	0.244	0.217	0.293	0.297
NAI	0.438	0.308	0.338	0.432	0.460	0.284	0.240	0.260	0.384	0.427	0.292	0.241	0.216	0.309	0.322
Δ IP	0.441	0.313	0.343	0.437	0.468	0.287	0.241	0.262	0.389	0.447	0.288	**0.236**	**0.213**	0.310	0.335
Δ Housing	0.402	0.277	0.300	0.386	0.406	0.279	0.234	0.249	**0.331**	**0.385**	0.295	0.249	0.219	0.300	0.298
VIX and NFCI	0.331	0.212	0.250	0.364	**0.351**	0.235	**0.203**	**0.228**	0.345	0.440	**0.254**	0.249	0.229	0.321	0.391
VIX and Δ Indpro	0.332	0.211	0.251	0.363	0.351	0.234	0.204	0.230	0.354	0.450	0.257	0.253	0.237	0.338	0.424
VIX and NAI	0.333	0.212	0.254	0.367	0.354	0.234	0.205	0.231	0.358	0.449	0.259	0.254	0.237	0.334	0.417
VIX and Δ Housing	0.330	0.213	0.254	0.369	0.359	0.237	0.209	0.234	0.343	0.439	0.260	0.261	0.242	0.333	0.409
Avg.	0.396	0.273	0.303	0.391	0.403	0.269	0.226	0.247	0.362	0.418	0.272	0.240	0.217	0.306	0.335
Competitor models															
GARCH	0.430	0.296	0.325	0.419	0.452	0.285	0.241	0.263	0.394	0.441	0.300	0.252	0.232	0.340	0.370
MS-GARCH-TVI	0.468	0.338	0.370	0.452	0.519	0.303	0.270	0.298	0.429	0.488	0.286	0.246	0.233	0.348	0.405
MS-GARCH-TVC	0.461	0.318	0.335	0.414	0.437	0.290	0.245	0.263	0.386	0.432	0.295	0.239	0.218	0.324	0.350
RealGARCH	0.237	**0.182**	**0.239**	0.380	0.409	0.256	0.208	0.229	0.358	0.408	0.331	0.261	0.229	**0.287**	**0.289**
HEAVY	0.272	0.223	0.326	0.591	0.759	0.262	0.228	0.273	0.498	0.593	0.339	0.305	0.313	0.535	0.642
HAR	0.234	0.189	0.254	**0.359**	0.385	0.243	0.212	0.238	0.374	0.430	0.340	0.257	0.230	0.371	0.470
HAR (lev.)	**0.226**	0.187	0.258	0.362	0.387	**0.232**	0.211	0.240	0.378	0.429	0.286	0.245	0.227	0.373	0.470
No-change	0.418	0.821	1.143	2.213	2.310	0.304	0.297	0.336	0.532	0.715	0.382	0.320	0.314	0.481	0.555

Notes: Numbers reported are the average out-of-sample QLIKE losses for each model for one-day- (1d), two-week- (2w), one-month- (1m), two-month- (2m) and three-month-ahead (3m) variance forecasts across three different volatility regimes; forecasts are issued at a day for which the daily realized volatility is below the empirical 25% quantile (low regime), between the 25% and 75% quantile (normal regime), or above the 75% quantile (high regime). Bold entries indicate the model with the lowest average QLIKE loss per regime and horizon. Blue-shaded numbers indicate that the respective model is included in the 90% model confidence set. The average forecast (avg.) is the mean forecast across all GARCH-MIDAS models employing one explanatory variable. The out-of-sample evaluation period spreads 2010:M1 to 2018:M4.

Table 1.8: Mincer-Zarnowitz R^2.

	Panel A: Full out-of-sample period					**Panel B: Volatility regimes**		
						low	**normal**	**high**
	1d	2w	1m	2m	3m	1m	1m	1m
GARCH-MIDAS								
RVol(22)	0.312	0.367	0.340	0.086	0.008	0.037	0.061	0.314
VIX	0.347	0.346	0.321	0.145	0.047	0.071	0.099	0.297
VRP	0.343	**0.404**	0.354	0.128	0.030	0.041	0.083	0.324
NFCI	0.295	0.375	0.354	0.146	0.062	0.030	0.073	0.341
NAI	0.294	0.373	0.352	0.143	0.062	0.025	0.071	0.339
Δ IP	0.296	0.374	0.348	0.124	0.029	0.017	0.065	0.341
Δ Housing	0.293	0.372	0.355	**0.168**	**0.102**	0.031	0.077	0.334
VIX and NFCI	**0.353**	0.359	0.333	0.147	0.050	0.072	0.100	0.302
VIX and NAI	0.348	0.349	0.323	0.146	0.048	0.067	0.099	0.297
VIX and Δ IP	0.348	0.347	0.321	0.145	0.047	0.070	0.099	0.296
VIX and Δ Housing	0.347	0.346	0.321	0.153	0.056	0.064	0.099	0.295
Avg.	0.322	0.380	0.357	0.149	0.057	0.036	0.078	0.341
Competitor models								
GARCH	0.288	0.373	0.353	0.138	0.051	0.027	0.068	0.343
MS-GARCH-TVI	0.316	0.357	0.288	0.118	0.016	0.005	0.015	0.339
MS-GARCH-TVC	0.311	0.390	0.368	0.142	0.052	0.030	0.066	**0.374**
RealGARCH	0.318	0.394	**0.377**	0.146	0.070	**0.076**	**0.112**	0.303
HEAVY	0.297	0.322	0.272	0.061	0.004	0.028	0.084	0.173
HAR	0.312	0.394	0.374	0.125	0.052	0.058	0.087	0.315
HAR (lev.)	0.342	0.392	0.366	0.122	0.053	0.056	0.088	0.303
No-change	0.254	0.227	0.189	0.060	0.020	0.046	0.044	0.088

Notes: We report coefficients of determination derived from MZ regressions. Bold entries indicate the models with the highest R^2 for a specific forecast horizon. The last three columns correspond to the forecast evaluation divided in three volatility regimes; forecasts are issued at a day for which the daily realized volatility is below the empirical 25% quantile (low regime), between the 25% and 75% quantile (normal regime), or above the 75% quantile (high regime). The out-of-sample evaluation period spreads 2010:M1 to 2018:M4.

1.6 Appendix

1.6.1 Proofs

Proof of Proposition 1.1. The proof follows directly by applying the mutual independence of $g_{i,t}$, τ_t and $Z_{i,t}$ and by noting that Assumption 1.3 implies $\mathbf{E}[\tau_t^2]/\mathbf{E}[\tau_t]^2 > 1$ if τ_t is non-constant. □

Proof of Proposition 1.2. First, note that under Assumptions 1.1, 1.2, and 1.3 the covariance $\mathbf{Cov}(\varepsilon_t^2, \varepsilon_{t-k}^2)$ exists for every $k \in \mathbb{N}$ and is time-invariant. In the proof, we use that τ_t and g_t are independent covariance stationary processes and that Z_t are *i.i.d.* innovations.

$$\begin{aligned}
\rho_k^{MG}(\varepsilon^2) &= \frac{\mathbf{Cov}(\varepsilon_t^2, \varepsilon_{t-k}^2)}{\sqrt{\mathbf{Var}(\varepsilon_t^2)}\sqrt{\mathbf{Var}(\varepsilon_{t-k}^2)}} \\
&= \frac{\mathbf{E}[\tau_t\tau_{t-k}]\mathbf{E}[g_tZ_t^2g_{t-k}Z_{t-k}^2] - \mathbf{E}[\tau_t]\mathbf{E}[\tau_{t-k}]}{\mathbf{Var}(\varepsilon_t^2)} \\
&= \frac{\mathbf{E}[\tau_t\tau_{t-k}]\mathbf{E}[g_tZ_t^2g_{t-k}Z_{t-k}^2] - \mathbf{E}[\tau_t\tau_{t-k}] + \mathbf{E}[\tau_t\tau_{t-k}] - \mathbf{E}[\tau_t]\mathbf{E}[\tau_{t-k}]}{\mathbf{Var}(\varepsilon_t^2)} \\
&= \frac{\mathbf{E}[\tau_t\tau_{t-k}] - \mathbf{E}[\tau_t]^2}{\mathbf{Var}(\varepsilon_t^2)} + \frac{\Big(\mathbf{E}[g_tZ_t^2g_{t-k}Z_{t-k}^2] - \mathbf{E}[g_t]\mathbf{E}[g_{t-k}]\Big)\mathbf{E}[\tau_t\tau_{t-k}]}{\mathbf{Var}(\varepsilon_t^2)} \\
&= \frac{\mathbf{Cov}(\tau_t, \tau_{t-k})}{\mathbf{Var}(\varepsilon_t^2)} + \frac{\mathbf{Cov}(g_tZ_t^2, g_{t-k}Z_{t-k}^2)(\mathbf{Cov}(\tau_t, \tau_{t-k}) + \mathbf{E}[\tau_t^2])}{\mathbf{Var}(\varepsilon_t^2)} \\
&= \rho_k^\tau \frac{\mathbf{Var}(\tau_t)}{\mathbf{Var}(\varepsilon_t^2)} + \rho_k^{GA}\frac{\left(\rho_k^\tau\mathbf{Var}(\tau_t) + \mathbf{E}[\tau_t]^2\right)\mathbf{Var}(g_tZ_t^2)}{\mathbf{Var}(\varepsilon_t^2)}
\end{aligned}$$

□

Proof of Proposition 1.3. Employing the assumptions used in the proof of Proposition 1.2 above, we conclude similarly:

$$\begin{aligned}
\rho_k^{MG}(\sigma^2) &= \frac{\mathbf{Cov}(\sigma_t^2, \sigma_{t-k}^2)}{\sqrt{\mathbf{Var}(\sigma_t^2)}\sqrt{\mathbf{Var}(\sigma_{t-k}^2)}} \\
&= \frac{\mathbf{E}[\tau_t\tau_{t-k}]\mathbf{E}[g_tg_{t-k}] - \mathbf{E}[\tau_t]\mathbf{E}[\tau_{t-k}]}{\mathbf{Var}(\sigma_t^2)} \\
&= \frac{\mathbf{E}[\tau_t\tau_{t-k}]\mathbf{E}[g_tg_{t-k}] - \mathbf{E}[\tau_t\tau_{t-k}] + \mathbf{E}[\tau_t\tau_{t-k}] - \mathbf{E}[\tau_t]\mathbf{E}[\tau_{t-k}]}{\mathbf{Var}(\sigma_t^2)} \\
&= \frac{\mathbf{E}[\tau_t\tau_{t-k}] - \mathbf{E}[\tau_t]^2}{\mathbf{Var}(\sigma_t^2)} + \frac{\left(\mathbf{E}[g_tg_{t-k}] - \mathbf{E}[g_t]\mathbf{E}[g_{t-k}]\right)\mathbf{E}[\tau_t\tau_{t-k}]}{\mathbf{Var}(\sigma_t^2)} \\
&= \frac{\mathbf{Cov}(\tau_t, \tau_{t-k})}{\mathbf{Var}(\sigma_t^2)} + \frac{\mathbf{Cov}(g_t, g_{t-k})(\mathbf{Cov}(\tau_t, \tau_{t-k}) + \mathbf{E}[\tau_t^2])}{\mathbf{Var}(\sigma_t^2)} \\
&= \rho_k^\tau \frac{\mathbf{Var}(\tau_t)}{\mathbf{Var}(\sigma_t^2)} + \rho_k^{g}\frac{\left(\rho_k^\tau\mathbf{Var}(\tau_t) + \mathbf{E}[\tau_t]^2\right)\mathbf{Var}(g_t)}{\mathbf{Var}(\sigma_t^2)}.
\end{aligned}$$

□

Proof of Proposition 1.4. Equation (1.4) follows directly from the mutual independence of $g_{i,t}$, τ_t, and $Z_{i,t}$. Next, Equation (1.4) is derived as

$$\begin{aligned}\mathbf{E}[g^2_{k,t+1|t}] &= \mathbf{E}\left[\left(1+(\alpha+\gamma/2+\beta)^{k-1}(g_{1,t+1|t}-1)\right)^2\right]\\ &= 1+2(\alpha+\gamma/2+\beta)^{k-1}\underbrace{(\mathbf{E}[g_{1,t+1|t}]-1)}_{=0}+(\alpha+\gamma/2+\beta)^{2(k-1)}(\mathbf{E}[g^2_{1,t+1|t}]-1)\\ &= 1+(\alpha+\gamma/2+\beta)^{2(k-1)}(\mathbf{E}[g^2_{1,t+1}]-1).\end{aligned}$$

In the last step, we use that $g_{1,t+1|t} = g_{1,t+1}$. Now, consider the first property: As $k \to \infty$, $\mathbf{E}[g^2_{k,t+1|t}]$ decreases monotonically towards one. Because the numerator decreases while the denominator is constant, R^2_k is decreasing in k. The limit follows readily from $\lim_{k\to\infty}\mathbf{E}[g^2_{k,t+1|t}] = 1$.

For deriving the second property, note that Equation (1.4) is a rational function of linear polynomials in $\mathbf{E}[\tau^2_{t+1}]$ with negative intercepts and positive gradients. By taking the first derivative, the signs of intercepts and gradients imply the rational function in $\mathbf{E}[\tau^2_{t+1}]$ to be strictly increasing. □

Proof of Lemma 1.1. Using Equation (1.3), we obtain

$$\begin{aligned}R^2_1 &= \frac{\mathbf{Var}(g_t\tau_t)}{\mathbf{Var}(\varepsilon^2_t)} = \frac{\mathbf{E}[g^2_t]\mathbf{E}[\tau^2_t]-\mathbf{E}[\tau_t]^2}{\mathbf{E}[g^2_t]\mathbf{E}[\tau^2_t]\kappa-\mathbf{E}[\tau_t]^2}\\ &= \frac{(1-(\alpha+\gamma/2+\beta)^2)\mathbf{E}[\tau^2_t]-(1-(\alpha+\gamma/2)^2\kappa-2(\alpha+\gamma/2)\beta-\beta^2)\mathbf{E}[\tau_t]^2}{(1-(\alpha+\gamma/2+\beta)^2)\mathbf{E}[\tau^2_t]\kappa-(1-(\alpha+\gamma/2)^2\kappa-2(\alpha+\gamma/2)\beta-\beta^2)\mathbf{E}[\tau_t]^2}.\end{aligned}$$

□

1.6.2 Additional tables

Table 1.9: Monte-Carlo parameter estimates of MS-GARCH-TVI.

	ω_1	ω_2	α	β	$p_{1,1}$	$p_{2,2}$
	Panel A: $Z_{n,i,t}$ normally distributed					
Monthly τ_t	0.029	0.050	0.057	0.910	0.997	0.995
	[0.024,0.034]	[0.038,0.067]	[0.053,0.062]	[0.902,0.917]	[0.992,0.999]	[0.982,0.998]
Daily τ_t	0.020	0.038	0.058	0.912	0.993	0.991
	[0.016,0.024]	[0.029,0.051]	[0.054,0.063]	[0.906,0.919]	[0.986,0.997]	[0.978,0.996]
	Panel B: $Z_{n,i,t}$ student-t distributed					
Monthly τ_t	0.028	0.066	0.052	0.914	0.993	0.980
	[0.021,0.035]	[0.050,0.088]	[0.045,0.058]	[0.904,0.925]	[0.984,0.997]	[0.941,0.994]
Daily τ_t	0.019	0.050	0.053	0.917	0.990	0.978
	[0.015,0.024]	[0.038,0.066]	[0.046,0.059]	[0.907,0.925]	[0.980,0.995]	[0.946,0.990]

Notes: The table reports the median MS-GARCH-TVI parameter estimates and in brackets the corresponding inter-quartile ranges across 2,000 Monte-Carlo simulations in which the true data-generating process is a GARCH-MIDAS model, see description of Table 1.1.

Table 1.10: Summary statistics of stock market returns and explanatory variables.

Variable	Freq.	Start	Obs.	Min.	Max.	Mean	Median	Sd.	Skew.	Kurt.
Stock market data										
S&P 500 returns	d	1971	11938	-22.90	10.96	0.03	0.04	1.06	-1.04	28.81
$\sqrt{\text{RV}}$	d	2000	4600	0.13	8.84	0.87	0.72	0.60	3.22	21.93
RVol(22)	d	1989	7390	0.23	5.54	0.95	0.80	0.56	2.97	17.46
Explanatory variables										
VIX	d	1990	7135	0.58	5.09	1.22	1.10	0.49	2.08	10.63
NFCI	w	1973	2470	-0.99	4.67	0.00	-0.33	1.00	1.94	6.53
NAI	m	1971	568	-5.16	2.76	-0.00	0.06	1.00	-1.21	6.96
Δ IP	m	1971	568	-4.43	2.38	0.18	0.22	0.72	-1.22	8.82
Δ Housing	m	1971	568	-30.67	25.67	-0.07	-0.19	8.03	-0.03	3.77

Notes: The table presents summary statistics for the different variables, whereby the column "Freq." indicates whether the data is observed on a daily (d), weekly (w) or monthly (m) frequency. The column "Start" indicates the year of the first observation for each variable. The data end in 2018:M4. The reported statistics include the number of observations (Obs.), the minimum (Min.) and maximum (Max.), the mean and median, the standard deviation (Sd.), the skewness (Skew.) and the kurtosis (Kurt.). We define $RVol(22)_{i,t} = \sqrt{1/22 \sum_{j=0}^{21} r_{i-j,t}^2}$. Changes in industrial production and housing starts are measured in month-over-month log differences, i.e. $\Delta X_t = 100 \cdot (\log(X_t) - \log(X_{t-1}))$.

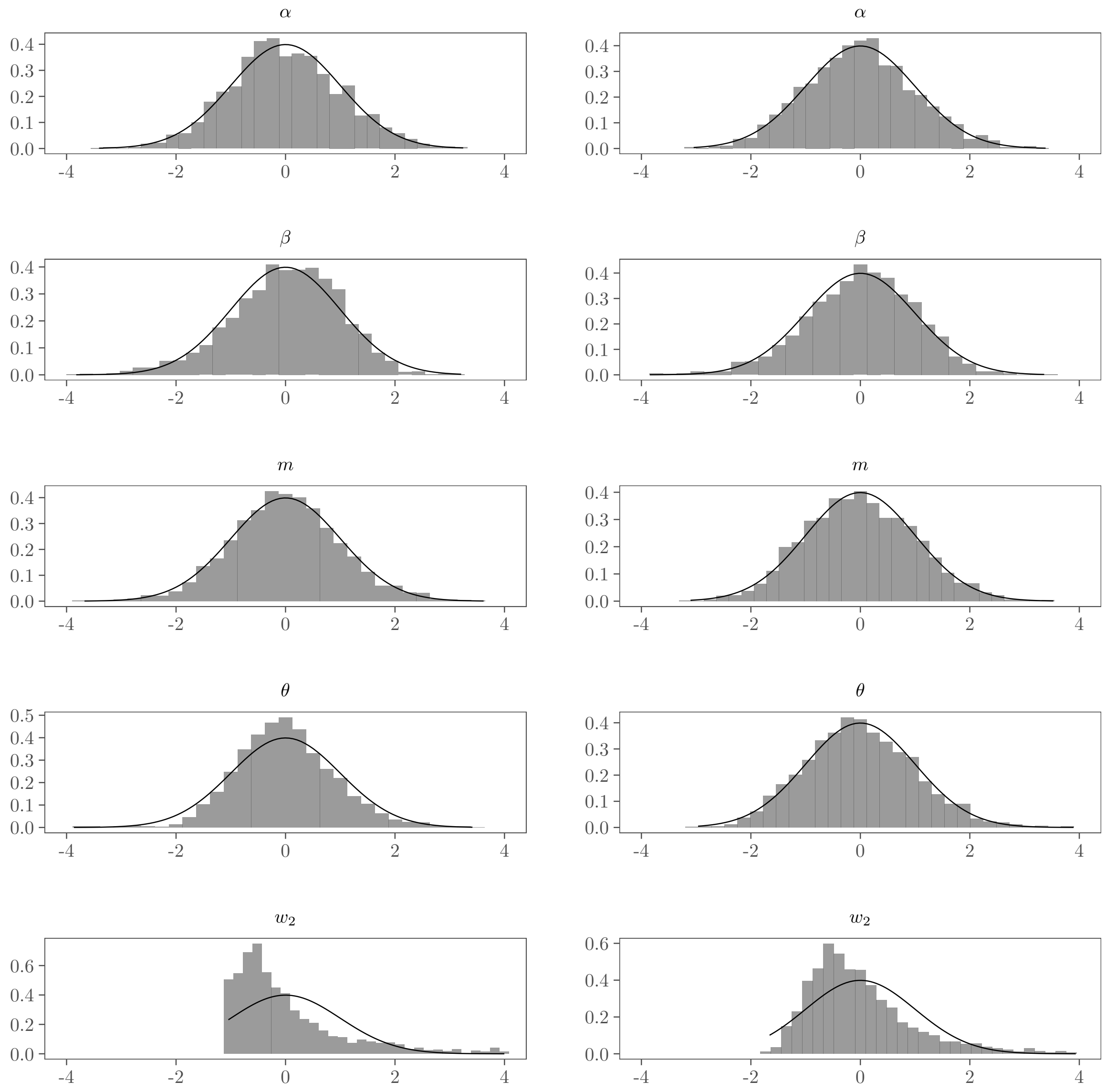
Table 1.11: Full sample estimates of competitor models.

Latent variables	Observables	Distribution
MS-GARCH-TVI (Haas et al., 2004)		
Regime 1: $\tilde{\sigma}^2_{1,t} = 0.008 + (0.018 + 0.092\mathbb{1}_{\{\varepsilon_{t-1}<0\}})\varepsilon^2_{t-1} + 0.918\tilde{\sigma}^2_{1,t-1}$ *Regime 2:* $\tilde{\sigma}^2_{2,t} = 0.183 + (0.018 + 0.092\mathbb{1}_{\{\varepsilon_{t-1}<0\}})\varepsilon^2_{t-1} + 0.918\tilde{\sigma}^2_{1,t-1}$ *Markov chain:* $X_t \in \{1,2\}$ with trans. prob. $p_{1,1} = 0.955$ and $p_{2,2} = 0.153$	$\varepsilon_t = \tilde{\sigma}_{X_t,t} Z_t$	$Z_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
MS-GARCH-TVC (Haas et al., 2004)		
Regime 1: $\tilde{\sigma}^2_{1,t} = 0.006 + (0.016 + 0.077\mathbb{1}_{\{\varepsilon_{t-1}<0\}})\varepsilon^2_{t-1} + 0.930\tilde{\sigma}^2_{1,t-1}$ *Regime 2:* $\tilde{\sigma}^2_{2,t} = 1.067 + 0.559\varepsilon^2_{t-1} + 0.439\tilde{\sigma}_{2,t-1}$ *Markov chain:* $X_t \in \{1,2\}$ with trans. prob. $p_{1,1} = 0.946$ and $p_{2,2} = 0.096$	$\varepsilon_t = \tilde{\sigma}_{X_t,t} Z_t$	$Z_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
Realized GARCH (Hansen et al., 2012)		
$\log \sigma^{RG}_t = 0.170 + 0.373 \log RV^{int}_{t-1} + 0.575 \log \sigma^{RG}_{t-1}$	$r_t - 0.018 = \sqrt{\sigma^{RG}_t} Z_t$ $\log RV^{int}_t = -0.472 + 1.056 \log \sigma^{RG}_t - 0.102 Z_t + 0.117 \left((Z_t)^2 - 1\right) + u_t$	$Z_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ $u_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.537)$
HEAVY (Shephard and Sheppard, 2010)		
$\sigma^{HVY}_t = 0.025 + 0.496 \cdot RV^{int}_{t-1} + 0.611 \cdot \sigma^{HVY}_{t-1}$ $\sigma^{RV^{int}}_t = 0.017 + 0.464 \cdot RV^{int}_{t-1} + 0.536 \cdot \sigma^{RV^{int}}_{t-1}$	$\varepsilon_t = \sqrt{\sigma^{HVY}_t} Z_t$ $RV^{int}_t = \sigma^{RV^{int}}_t Z^2_{RV^{int},t}$	$Z_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ $Z_{RV^{int},t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
HAR (Corsi, 2009)		
	$\log(RV_{t+1}) = -0.117 + 0.353 \log RV_t + 0.380 \log\left(\frac{RV_{t-4:t}}{5}\right) + 0.211 \log\left(\frac{RV_{t-21:t}}{22}\right) + \zeta_t$	$\mathbf{E}[\zeta_{t,k} \mid \mathcal{F}_{t-1,k}] = 0$
HAR (lev.) (Corsi and Reno, 2012)		
	$\log(RV_{t+1}) = -0.142 + 0.260 \log RV_t + 0.365 \log\left(\frac{RV_{t-4:t}}{5}\right) + 0.290 \log\left(\frac{RV_{t-21:t}}{22}\right)$ $-0.078 r_t - 0.208 \frac{r_{t-4:t}}{5} - 0.155 \frac{r_{t-21:t}}{22} + \zeta^{lev}_t$	$\mathbf{E}[\zeta^{lev}_{t,k} \mid \mathcal{F}_{t-1,k}] = 0$

Notes: Benchmark models introduced in Section 4.4.1 and their corresponding full-sample estimates. r_t, RV^{int}_t, and RV_t denote the return, the intraday realized variance and its overnight return-augmented close-close equivalent. All benchmark models assume $I_t = 1$, i.e. they are defined on a daily frequency. Distributional assumptions are those employed for estimation, e.g. normal innovations for QML estimation.

1.6.3 Additional figures

Figure 1.7: Histograms of standardized GARCH-MIDAS parameter estimates.



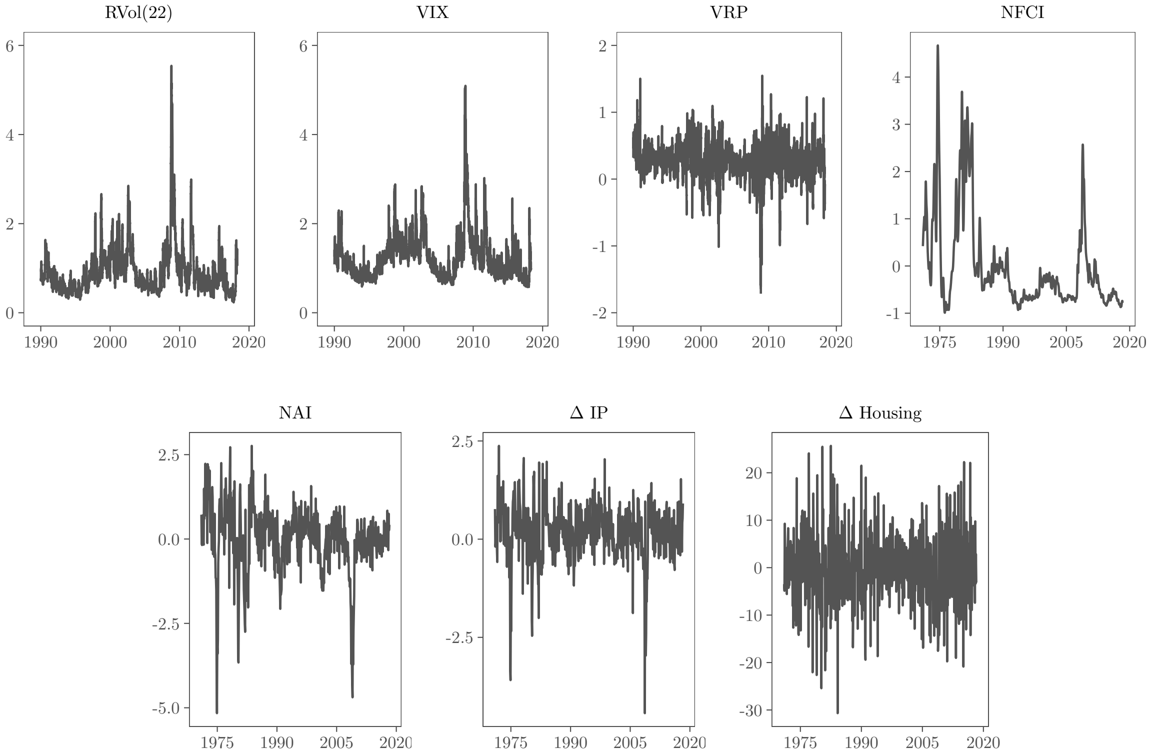
(a) Monthly τ_t

(b) Daily τ_t

Notes: Standardized empirical distributions of parameter estimates across 2,000 simulations are reported. On the left, the underlying data is generated by a GARCH-MIDAS model with monthly varying τ_t, on the right with daily varying τ_t, see Section 1.4 for further details. The standard normal distribution is depicted in black.

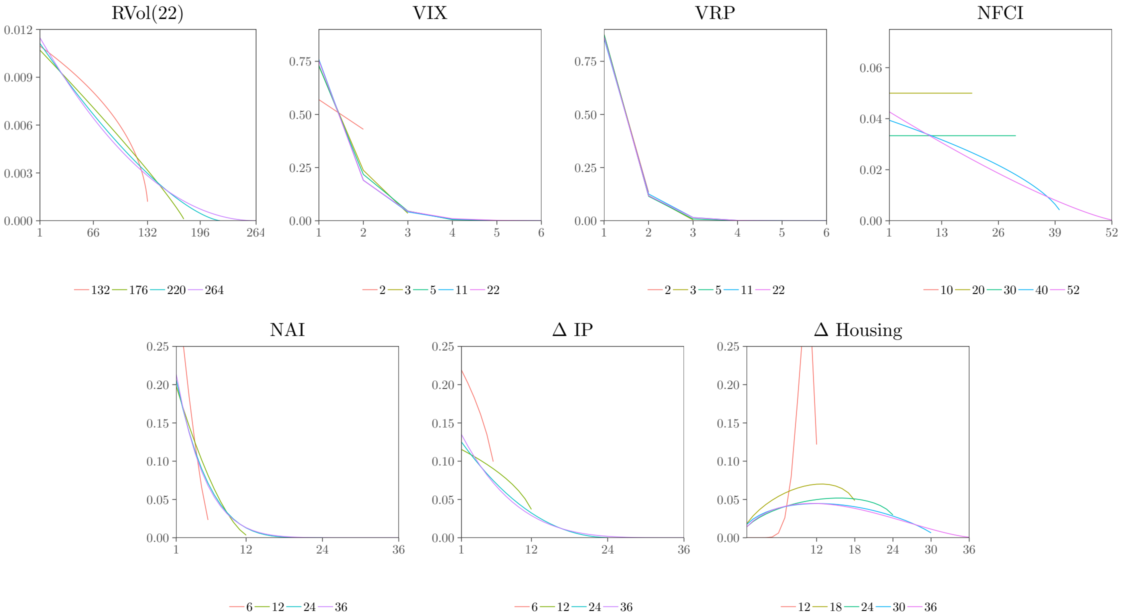
.

Figure 1.8: Time series of explanatory variables.



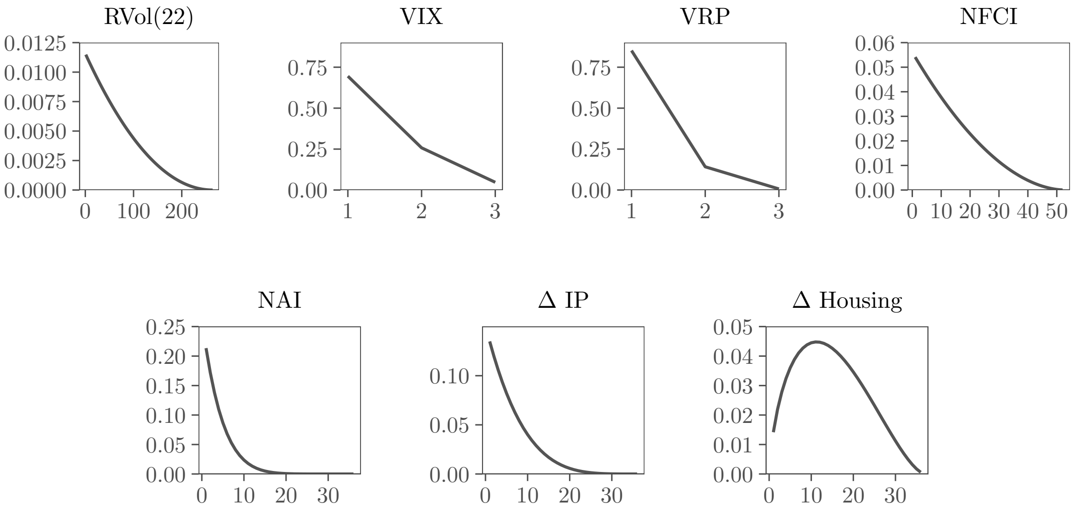
Notes: Daily financial data for the 1990:M1 to 2018:M4 period and macroeconomic data for the 1971:M1 to 2018:M4 period. See Section 4.1 for definitions and Table 1.10 for descriptive statistics of those variables.

Figure 1.9: Selected weighting schemes for different lag lengths.



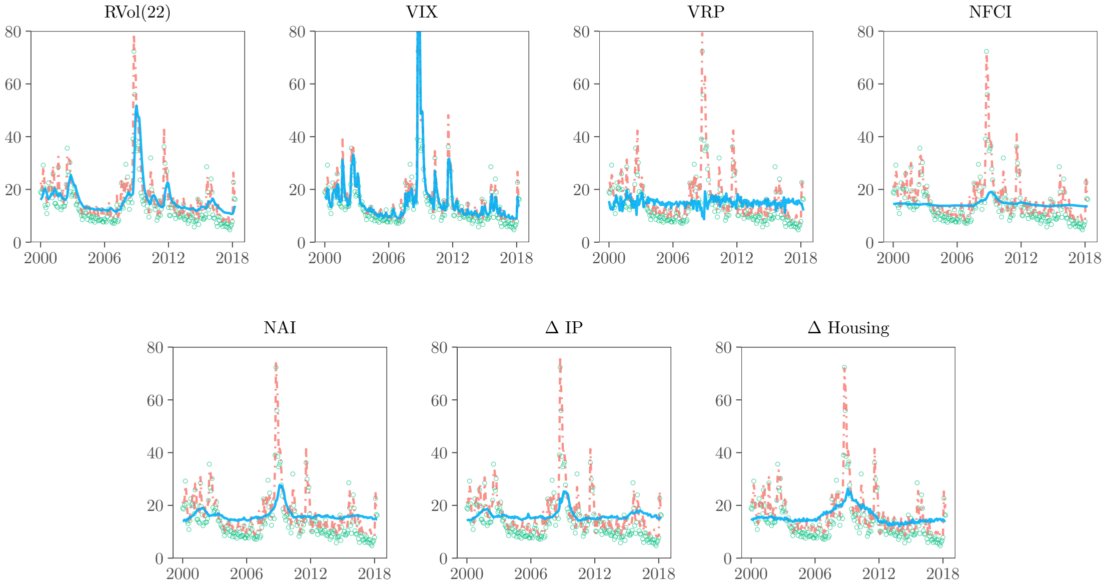
Notes: We depict selected weighting schemes that are implied by full-sample estimates for additional lag lengths K compared to those discussed in our empirical analysis.

Figure 1.10: Weighting schemes for different explanatory variables.



Notes: For each explanatory variable, the estimated Beta weighting scheme (see Equation (9)) based on full sample estimates is depicted. For all variables except housing starts, we impose the restriction $w_1 = 1$. The corresponding parameters are reported in Table 1.4.

Figure 1.11: Estimated monthly conditional volatility components.



Notes: The figure shows the monthly long-run volatility components $\sqrt{\tau_M}$ (blue, solid) and the monthly conditional volatilities $\sqrt{g_M \tau_M}$ (red, dot-dashed) for all GARCH-MIDAS models. To ensure comparability across the seven models, all figures cover the 2000:M1 to 2018:M4 period. Circles correspond to realized volatilities. Volatility is measured on an annualized scale.

1.6.4 Simulations: Violation of Assumption 3

In the following we present the results of two additional simulations. The simulations cover scenarios in which Assumption 1.3 is violated. In this section, we consider daily explanatory variables (i.e. we set $I_t = 1$) because empirically a violation of Assumption 1.3 is more likely to occur for daily explanatory variables than for low-frequency explanatory variables. Both simulations show that even if Assumption 1.3 is violated, our theoretical results still apply.

First, we consider a daily explanatory variable, X_t, that is correlated with the daily innovations Z_t.[29] Recall that in our simulation the daily innovations are given by

$$Z_t = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_{n,t},$$

i.e. $Z_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. As before, we model X_t as an AR(1) process

$$X_t = \phi X_{t-1} + \xi_t$$

but the innovation is now given by

$$\xi_t / \sigma_\xi = \rho_{\xi,Z} Z_t + \sqrt{1 - \rho_{\xi,Z}^2}\tilde{\xi}_t,$$

where $\tilde{\xi}_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, independent of Z_t and $\rho_{\xi,Z} \in [-1,1]$. In this setting, the correlation between the daily innovations Z_t and ξ_t is $\rho_{\xi,Z}$. We set $\rho_{\xi,Z} = -0.8$. The negative correlation between innovations to returns and innovations to X_t mimic the fact that changes in returns and daily measures of risk (such as the VIX index) are typically negatively correlated. Under our choice of $\phi = 0.98$, the contemporaneous correlation between Z_t and X_t is -0.16. Z_t is also correlated with future X_t but uncorrelated with past X_t.

In Table 1.12, Panel A shows that on average the QML estimates are still close to the true parameter values and the asymptotic standard errors are accurate. Most importantly, Panel A of Figure 1.12 illustrates that our results regarding the R^2 of a MZ regression still hold when X_t and Z_t are correlated. Panel A of Table 1.13 shows the corresponding MCS inclusion rates. Clearly, the correctly specified GARCH-MIDAS model with $K = 264$ and the GARCH-MIDAS with misspecified lag-length still do very well. In contrast, for forecast horizons of up to two

[29] Since $I_t = 1$, we can drop the index i.

months the forecast performance of the MS-GARCH-TVI appears to deteriorate considerably.

Second, we consider the GARCH-MIDAS-RV model, i.e. we choose

$$X_t = RVol(22)_t = \sqrt{\frac{1}{22}\sum_{j=0}^{21} r_{t-j}^2}.$$

This choice corresponds to the GARCH-MIDAS-RV specification that is estimated in the empirical application in Section 4. Again, Z_t is correlated with the contemporaneous and future X_t but uncorrelated with lagged X_t. The results for this specification are presented in Panels B of Table 1.12, Figure 1.12 and Table 1.13. Again, our previous findings regarding the MZ R^2 and the MCS inclusion rates are confirmed.

Table 1.12: Monte-Carlo parameter estimates: X_t and Z_t dependent.

	α	β	m	θ	w_2	$\kappa - 3$
Panel A: innovations to X_t correlated with Z_t						
GARCH-MIDAS (264)	0.000	-0.003	-0.001	0.008	0.890	-0.008
		{0.008}	{0.014}	{0.064}	{0.075}	{5.675}
		(0.008)	(0.014)	(0.063)	(0.075)	(7.741)
GARCH-MIDAS (66)	0.000	-0.003	0.002	-0.055	-3.185	-0.006
GARCH	0.003	0.003	0.034	—	—	0.017
Panel B: X_t given by *RVol(22)*$_t$						
GARCH-MIDAS (264)	-0.043	-0.034	0.370	-0.533	0.629	0.025
		{0.013}	{0.098}	{0.599}	{0.589}	{2.432}
		(0.013)	(0.079)	(0.329)	(0.321)	(4.555)
GARCH-MIDAS (66)	-0.045	-0.026	1.067	-1.230	1.823	0.032
GARCH	-0.052	0.087	1.373	—	—	0.048

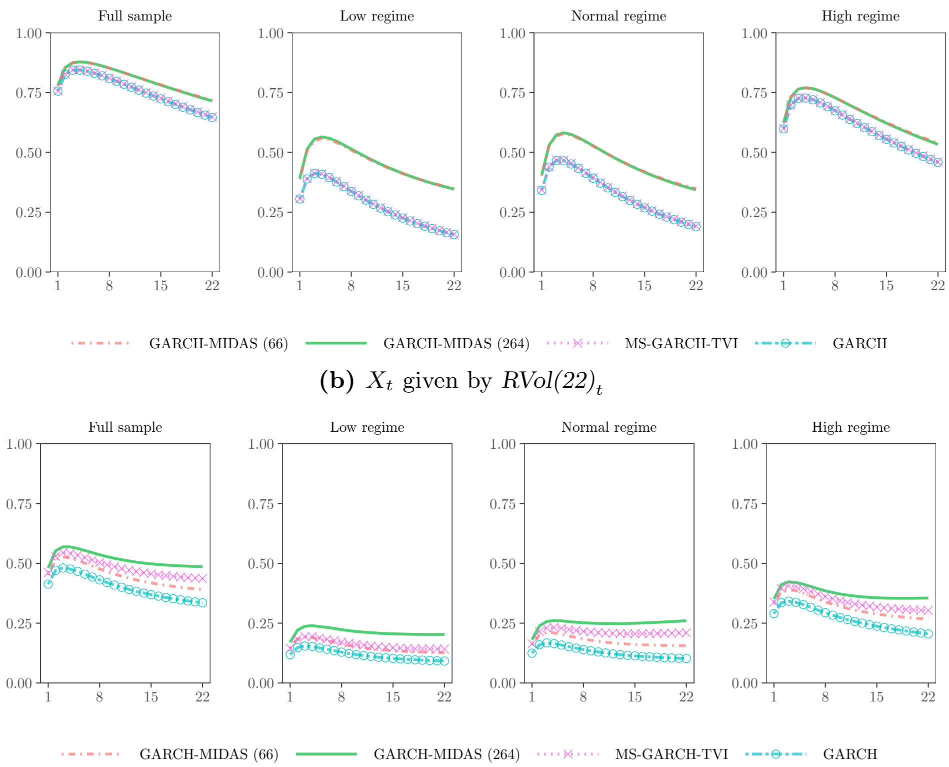
Notes: Modified version of Panel A in Table 1.1 for the case of a daily varying long-term component but Assumption 1.3 being violated. In Panel A, the true parameters are the same as in Table 1.1. However, the innovations ξ_t in the AR(1) process of X_t are correlated with Z_t, $\xi_t/\sigma_\xi = \rho_{\xi,Z} Z_t + \sqrt{1-\rho_{\xi,Z}^2}\tilde{\xi}_t$, $\tilde{\xi}_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. In Panel B, Assumption 3 is violated by employing a rolling window of past realized volatilities as a covariate, i.e. $X_t = RVol(22)_t = \sqrt{\frac{1}{22}\sum_{j=0}^{21} r_{t-j}^2}$. In this case, the GARCH-MIDAS parameters are given by $\mu = 0$, $\alpha = 0.1$, $\beta = 0.8$, $K = 264$, $m = -1$, $\theta = 1.6$, and $w_2 = 2.1$.

Table 1.13: Model confidence set inclusion rates: X_t and Z_t dependent.

	1d	2w	1m	2m	3m
Panel A: innovations to X_t correlated with Z_t					
GARCH-MIDAS (264)	0.953	0.896	0.867	0.802	0.755
GARCH-MIDAS (66)	0.848	0.786	0.832	0.882	0.874
MS-GARCH-TVI	0.362	0.100	0.135	0.471	0.757
GARCH	0.259	0.038	0.048	0.251	0.496
Panel B: X_t given by $\mathit{RVol(22)}_t$					
GARCH-MIDAS (264)	0.932	0.892	0.887	0.878	0.857
GARCH-MIDAS (66)	0.371	0.140	0.097	0.197	0.301
MS-GARCH-TVI	0.743	0.654	0.640	0.757	0.827
GARCH	0.152	0.048	0.046	0.098	0.138

Notes: Modified version of the upper panel of Table 1.2 for two cases in which X_t depends on (past values of) Z_t. See notes of Table 1.12 for a detailed description of these two scenarios.

Figure 1.12: MZ $R^2_{1:k}$—evaluation based on $RV_{1:k,t+1}$—X_t and Z_t dependent.



(a) innovations to X_t correlated with Z_t

(b) X_t given by $\mathit{RVol(22)}_t$

Notes: Modified version of Figure 1.5 for two scenarios in which X_t depends on (past values of) Z_t. See notes of Table 1.12 for a detailed description of these two scenarios.

1.6.5 Simulation with diffusion limit

In this section, we present simulation results for a situation in which the short-term discrete-time GARCH component (Equation (1.2)) has been replaced by its diffusion limit (see Nelson, 1990). In accordance with Andersen and Bollerslev (1998, pp. 894–895 and footnote 18 in the main text), we simulate the continuous-time data generating process using an Euler discretization scheme:

$$\varepsilon_{s+\Delta,t} = \log P_{s+\Delta,t} - \log P_{s,t} = \sqrt{\tau_t \tilde{g}_{s+\Delta,t}\Delta} W_{P,s,t}$$

with

$$\tilde{g}_{s+\Delta,t} = \tilde{\theta}\Delta + \tilde{g}_{s,t}\left(1 - \tilde{\theta}\Delta + \sqrt{2\tilde{\theta}\tilde{\lambda}\Delta} W_{\tilde{g},s,t}\right),$$

where $W_{P,s,t}$ and $W_{\tilde{g},s,t}$ are independent standard normal variables and the unit-variance GARCH-consistent parameters are given by

$$\tilde{\theta} = -\log(\alpha + \beta)$$

and

$$\begin{aligned}\tilde{\lambda} = 2\log(\alpha+\beta)^2 \cdot \Bigg\{ &\left(\left(1-(\alpha+\beta)^2\right)\cdot(1-\beta)^2\right)\cdot\alpha^{-1}\cdot(1-\beta\cdot(\alpha+\beta))^{-1}\Big) \\ &+ 6\cdot\log(\alpha+\beta) + 2\cdot\log(\alpha+\beta)^2 + 4\cdot(1-\alpha-\beta)\Bigg\}^{-1}.\end{aligned}$$

We choose Δ such that we obtain 20 price changes per five-minute interval.

Tables 1.14 and 1.15 are the equivalent of Tables 1.1 and 1.2. Figures 1.13 and 1.14 are the equivalent of Figures 1.4 and 1.5.

As expected, the parameter estimates in Table 1.14 are close to the ones in Table 1.1. Only in the case of a monthly τ_t do we observe an increase in bias for w_2. Moreover, we note that the excess kurtosis is considerably higher, even in comparison to our results regarding Student's t distributed intraday returns. Figure 1.13 makes it clear that we observe the same effect as in Figure 4. The same holds for Figure 1.14 and the corresponding Figure 1.5 in the main text. Likewise, the MCS inclusion rates reported in Table 1.15 confirm the overall results of Table 1.2 qualitatively. However, the MS-GARCH-TVI and GARCH models are less often excluded from the MCS.

Table 1.14: Monte-Carlo parameter estimates with GARCH diffusion.

		α	β	m	θ	w_2	$\kappa-3$
Monthly τ_t	GARCH-MIDAS (36)	-0.000	-0.007	-0.010	0.037	3.905	0.404
	GARCH-MIDAS (12)	-0.000	-0.006	-0.009	-0.029	0.396	0.406
	GARCH-MIDAS (36, $\tilde{X}$)	-0.000	-0.006	-0.009	-0.001	1.476	0.406
	GARCH-MIDAS (12, $\tilde{X}$)	-0.000	-0.005	-0.008	-0.076	-0.818	0.407
	GARCH	-0.000	0.001	0.005	—	—	0.421
Daily τ_t	M-GARCH (264)	-0.000	-0.006	-0.005	0.010	1.008	0.410
	GARCH-MIDAS (66)	-0.000	-0.005	-0.003	-0.050	-3.281	0.412
	GARCH-MIDAS (264, $\tilde{X}$)	-0.000	-0.006	-0.005	0.003	0.369	0.411
	GARCH-MIDAS (66, $\tilde{X}$)	0.000	-0.005	-0.002	-0.061	-3.448	0.414
	GARCH	0.003	0.001	0.030	—	—	0.442

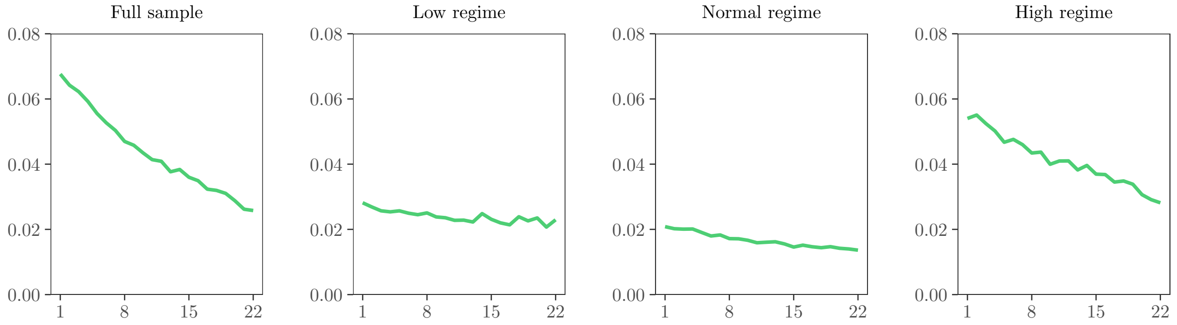
Notes: Modified version of the upper panel of Table 1.1. The only difference is that the short-term GARCH component is replaced by a consistent diffusion limit.

Table 1.15: Model confidence set inclusion rates with GARCH diffusion.

		1d	2w	1m	2m	3m
Monthly τ_t	GARCH-MIDAS (36)	0.919	0.864	0.845	0.823	0.811
	GARCH-MIDAS (12)	0.918	0.873	0.854	0.846	0.837
	GARCH-MIDAS (36, $\tilde{X}$)	0.874	0.784	0.757	0.742	0.720
	M-GARCH (12, $\tilde{X}$)	0.852	0.784	0.746	0.734	0.715
	MS-GARCH-TVI	0.875	0.842	0.815	0.775	0.744
	GARCH	0.771	0.621	0.571	0.495	0.477
Daily τ_t	GARCH-MIDAS (264)	0.966	0.944	0.927	0.860	0.809
	GARCH-MIDAS (66)	0.935	0.915	0.916	0.907	0.880
	GARCH-MIDAS (264, $\tilde{X}$)	0.932	0.875	0.833	0.801	0.764
	GARCH-MIDAS (66, $\tilde{X}$)	0.905	0.860	0.841	0.848	0.855
	MS-GARCH-TVI	0.741	0.615	0.561	0.699	0.839
	GARCH	0.676	0.478	0.412	0.497	0.627

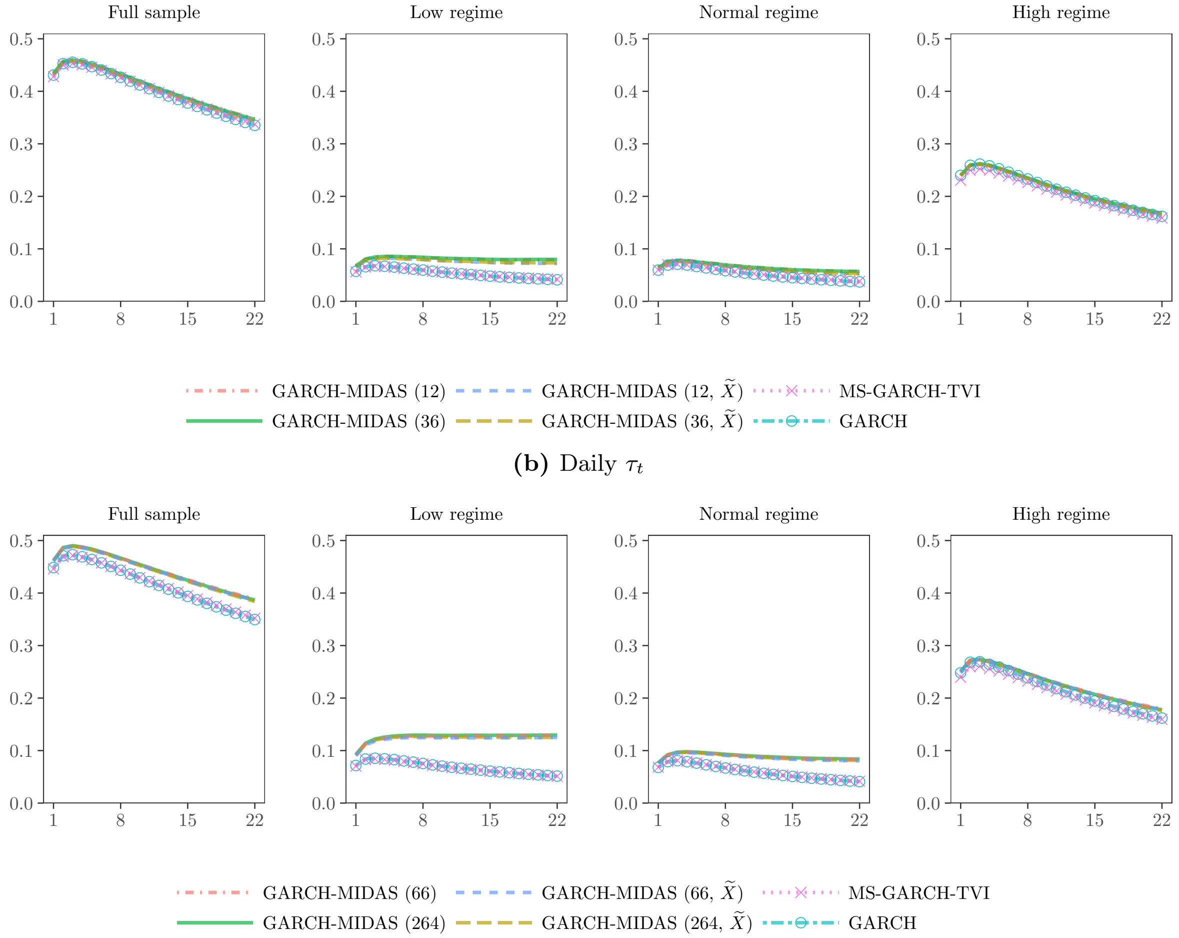
Notes: Modified version of the upper panel of Table 1.2. The only difference is that the short-term GARCH component is replaced by a consistent diffusion limit.

Figure 1.13: MZ R^2—monthly τ_t—evaluation based on $\varepsilon^2_{k,t+1}$ (with diffusion).



Notes: Modified version of Figure 1.4. The only difference is that the short-term GARCH component is replaced by a consistent diffusion limit.

Figure 1.14: MZ R^2—monthly and daily τ_t—evaluation based on $RV_{1:k,t+1}$ (with diffusion).



(a) Monthly τ_t

(b) Daily τ_t

Notes: Modified version of Figure 1.5. The only difference is that the short-term GARCH component is replaced by a consistent diffusion limit.

1.6.6 Additional details about data

In this section, we provide detailed information on the data sources as well as on the data vintages that have been used. Whenever possible, we use real-time vintage data sets as available in ALFRED.[30] For downloading the respective data sources, we have written the R-package (Kleen, 2017).[31] We make use of the following time series:

- Realized volatility based on five-minute intraday returns which are provided by the Realized Library of the Oxford-Man Institute of Quantitative Finance (Heber et al., 2009).

  http://realized.oxford-man.ox.ac.uk/data/download/

- The Cboe Volatility Index (VIX) as a measure of option-implied volatility of S&P 500 returns (published by the Chicago Board Options Exchange).

  http://www.cboe.com/micro/vix/historical.aspx

- The Chicago Fed's National Financial Conditions Index (NFCI), measuring the risk, liquidity and leverage of money markets, debt and equity markets, and the traditional and shadow banking system. The NFCI takes positive/negative values whenever financial conditions are tighter/looser than on average.

  https://alfred.stlouisfed.org/series?seid=NFCI

- The Chicago Fed National Activity Index (NAI) is a weighted average of 85 filtered and standardized economic indicators. Whereas positive NAI values indicate an expanding US-economy above its historical trend rate, negative values indicate the opposite.

  https://alfred.stlouisfed.org/series?seid=CFNAI

- Industrial Production Index (IP), which is released by the Board of Governors of the Federal Reserve System.

  https://alfred.stlouisfed.org/series?seid=INDPRO

- New Privately Owned Housing Units Started (HOUST), which is published by the U.S. Bureau of the Census.

  https://alfred.stlouisfed.org/series?seid=HOUST

For the macroeconomic variables, we report the real-time data availability in Table 1.16.

[30] https://alfred.stlouisfed.org

[31] https://cran.r-project.org/package=alfred

Table 1.16: Real-time data availability.

Variable	Frequency	ALFRED ID	First Vintage Release
NFCI	weekly	NFCI	2011-05-25
NAI	monthly	CFNAI	2011-05-23
Industrial production	monthly	INDPRO	1973-12-14
Housing starts	monthly	HOUST	1973-12-18

Note: For each macroeconomic variable, we report the real-time data availability in the ALFRED data base.

1.6.7 Description of benchmark models

For the empirical implementation, we use the statistical computing environment R (R Core Team, 2018, *R: A Language and Environment for Statistical Computing.* https://www.r-project.org/). In the following, we present some details regarding the specification and estimation of the different models. For all benchmark models we have that $I_t = 1$ and, hence, the index i can be dropped.

- Two **Markov-Switching GARCH** models (MS-GARCH-TVI and MS-GARCH-TVC): Our specification follows (Haas, Mittnik, and Paolella, 2004). Returns are decomposed as $\varepsilon_t = \tilde{\sigma}_{X_t,t} Z_t$, where $\{X_t\}$ is a Markov chain with a finite state space $S = \{1, 2\}$. The conditional variance in state $X_t = k$ is given by

$$\tilde{\sigma}^2_{k,t} = \omega_k + (\alpha_k + \gamma_k \mathbb{1}_{\{Z_{t-1}<0\}})\varepsilon^2_{t-1} + \beta_k \tilde{\sigma}^2_{k,t-1}.$$

  We employ two different specifications which nest the baseline GJR-GARCH model:

  1. An MS-GARCH called MS-GARCH-TVI (time-varying intercept) in which only the intercept is driven by the Markov chain while the ARCH/GARCH parameters are the same in both equations. In the simulations we set $\gamma_k = 0$.
  2. An MS-GARCH called MS-GARCH-TVC (time-varying coefficients) which models one regime as a GJR-GARCH and another regime as a standard GARCH(1,1), i.e. $\gamma_2 = 0$.[32]

  For estimation, we use the R-package *MSGARCH*, v2.3, by Ardia et al. (2019). In both specifications we assume the innovations to be normally distributed which was numerically the most stable.

- As a generalization of the GARCH model, we employ the **Realized GARCH** model

[32] Modeling both regimes as a GJR-GARCH turned out to be numerically unstable.

(Hansen, Huang, and Shek, 2012). Here, the conditional variance of the returns $r_t - \mu^{RG} = \sqrt{\sigma_t^{RG}} Z_t^{RG}$, $Z_t^{RG} \overset{i.i.d.}{\sim} \mathcal{D}(0,1)$ at day t is modeled as

$$\log \sigma_t^{RG} = \omega^{RG} + \alpha^{RG} \log RV_{t-1}^{int} + \beta^{RG} \log \sigma_{t-1}^{RG}$$

and the realized measure RV_t^{int} as

$$\log RV_t^{int} = \xi^{RG} + \delta^{RG} \log \sigma_t^{RG} + \eta_1^{RG} Z_t^{RG} + \eta_2^{RG} \left(\left(Z_t^{RG} \right)^2 - 1 \right) + u_t^{RG}$$

with $u_t^{RG} \overset{i.i.d.}{\sim} \mathcal{N}(0, \lambda^{RG})$. The innovations Z_t^{RG} and u_t^{RG} are independent. The estimation of the Realized GARCH model and the forecast computation by simulation is carried out using the R-package *rugarch* (Ghalanos, 2018).

- The **HEAVY** model by Shephard and Sheppard (2010) is a joint model of returns and some realized measure. We use the intraday realized variance, RV_t^{int}, as the realized measure. The conditional variance equation of daily returns is given by

$$\mathbf{Var}(\varepsilon_t^2 | \mathcal{F}_{t-1}) =: \sigma_t^{HVY} = \omega_1^{HVY} + \alpha_1^{HVY} RV_{t-1}^{int} + \beta_1^{HVY} \sigma_{t-1}^{HVY}$$

and the realized measure equation by

$$\mathbf{E}[RV_t^{int} | \mathcal{F}_{t-1}] =: \sigma_t^{RV^{int}} = \omega_2^{HVY} + \alpha_2^{HVY} RV_{t-1}^{int} + \beta_2^{HVY} \sigma_{t-1}^{RV^{int}}.$$

We assume $\omega_1^{HVY}, \omega_2^{HVY}, \alpha_1^{HVY}, \alpha_2^{HVY}, \beta_2^{HVY} \geq 0$; $\beta_1^{HVY} \in [0,1)$; and $\alpha_2^{HVY} + \beta_2^{HVY} \in [0,1)$. The estimation is carried out by QML estimation. Note that both dynamic equations can be estimated separately. Often, the conditional variance equation is estimated to be unit-root. We compute iterative multi-step-ahead forecasts, see Shephard and Sheppard (2010, Equation (11), p. 205).

- We also consider a **HAR** specification that models realized variances directly (see Corsi, 2009). We specify the HAR model in terms of the log of the realized variances. The model for forecasting the k-period cumulative variance is given by

$$\log \left(\frac{RV_{t+1:t+k}}{k} \right) = b_0 + b_1 \log RV_t + b_2 \log \left(\frac{RV_{t-4:t}}{5} \right) + b_3 \log \left(\frac{RV_{t-21:t}}{22} \right) + \zeta_{t,k}$$

with $RV_{t+1:t+k} = \sum_{i=1}^{k} RV_{t+i}$. The HAR model is estimated by OLS. Realized variance forecasts are obtained as follows:

$$RV_{t+1:t+k|t} = k{\cdot}\exp\left(b_0 + b_1 \log RV_t + b_2 \log\left(\frac{RV_{t-4:t}}{5}\right) + b_3 \log\left(\frac{RV_{t-21:t}}{22}\right) + \frac{1}{2}\mathbf{Var}(\zeta_{t,k})\right),$$

assuming the residuals $\zeta_{t,k}$ to be normally distributed.

- **HAR with leverage** (Corsi and Renò, 2012):

$$\log\left(\frac{RV_{t+1:t+k}}{k}\right) = b_0^{lev} + b_1^{lev} \log RV_t + b_2^{lev} \log\left(\frac{RV_{t-4:t}}{5}\right) + b_3^{lev} \log\left(\frac{RV_{t-21:t}}{22}\right)$$
$$+ b_4^{lev} r_t + b_5^{lev} \times \frac{r_{t-4:t}}{5} + b_6^{lev} \times \frac{r_{t-21:t}}{22} + \zeta_{t,k}^{lev}$$

As in the case of the HAR model without leverage effect, we assume the residuals $\zeta_{t,k}^{lev}$ to be normally distributed in order to get closed-form expressions for the respective forecasts.

- The estimation of the **GARCH-MIDAS** models (see Section 1.2) has been carried out using QMLE, see Engle, Ghysels, and Sohn (2013), and can be replicated using the R-package *mfGARCH*, v0.1.8, by Kleen (2018).[33]

[33] https://cran.r-project.org/package=mfGARCH

1.6.8 95% model confidence sets

As a robustness check, the following Tables 1.17 and 1.18 replicate Tables 1.6 and 1.7 for a confidence level of 95% instead of 90%.

Table 1.17: QLIKE losses and 95% model confidence sets: full out-of-sample period.

	Full sample				
	1d	2w	1m	2m	3m
RVol(22)	0.306	0.246	0.271	0.387	0.428
VIX	0.275	0.215	0.240	0.359	0.414
VRP	0.291	0.227	0.260	0.384	0.430
NFCI	0.324	0.248	0.264	0.363	0.393
NAI	0.343	0.266	0.283	0.391	0.424
Δ IP	0.345	0.267	0.285	0.395	0.438
Δ Housing	0.328	0.252	0.264	**0.347**	**0.380**
VIX and NFCI	0.274	0.213	0.236	0.349	0.399
VIX and NAI	0.275	0.215	0.241	0.358	0.409
VIX and Δ IP	0.274	0.214	0.239	0.355	0.409
VIX and Δ Housing	0.275	0.218	0.243	0.351	0.405
Avg.	0.317	0.246	0.264	0.364	0.400
GARCH	0.342	0.263	0.282	0.395	0.434
MS-GARCH-TVI	0.362	0.292	0.315	0.426	0.488
MS-GARCH-TVC	0.355	0.271	0.283	0.387	0.421
RealGARCH	0.260	**0.206**	**0.233**	0.356	0.390
HEAVY	0.277	0.238	0.299	0.539	0.662
HAR	0.254	0.210	0.243	0.368	0.419
HAR (lev.)	**0.238**	0.207	0.245	0.371	0.419
No-change	0.358	0.498	0.636	1.157	1.292

Notes: See Table 1.6 but for a confidence level of 95% instead of 90%.

Table 1.18: QLIKE losses and 95% model confidence sets: low/normal/high-volatility regimes.

	Low-volatility regime					Normal-volatility regime					High-volatility regime				
	1d	2w	1m	2m	3m	1d	2w	1m	2m	3m	1d	2w	1m	2m	3m
RVol(22)	0.364	0.264	0.305	0.399	0.409	0.271	0.232	0.260	0.400	0.463	0.273	0.241	0.217	0.313	0.365
VIX	0.332	0.210	0.250	0.367	0.354	0.233	0.204	0.231	0.355	0.454	0.262	0.259	0.243	0.347	0.437
VRP	0.349	0.237	0.288	0.405	0.424	0.252	0.215	0.245	0.375	0.440	0.266	0.238	0.237	0.360	0.414
NFCI	0.400	0.274	0.304	0.389	0.402	0.272	0.228	0.248	0.364	0.417	0.292	0.244	0.217	0.293	0.297
NAI	0.438	0.308	0.338	0.432	0.460	0.284	0.240	0.260	0.384	0.427	0.292	0.241	0.216	0.309	0.322
Δ IP	0.441	0.313	0.343	0.437	0.468	0.287	0.241	0.262	0.389	0.447	0.288	**0.236**	**0.213**	0.310	0.335
Δ Housing	0.402	0.277	0.300	0.386	0.406	0.279	0.234	0.249	**0.331**	**0.385**	0.295	0.249	0.219	0.300	0.298
VIX and NFCI	0.331	0.212	0.250	0.364	**0.351**	0.235	**0.203**	**0.228**	0.345	0.440	**0.254**	0.249	0.229	0.321	0.391
VIX and Δ Indpro	0.332	0.211	0.251	0.363	0.351	0.234	0.204	0.230	0.354	0.450	0.257	0.253	0.237	0.338	0.424
VIX and NAI	0.333	0.212	0.254	0.367	0.354	0.234	0.205	0.231	0.358	0.449	0.259	0.254	0.237	0.334	0.417
VIX and Δ Housing	0.330	0.213	0.254	0.369	0.359	0.237	0.209	0.234	0.343	0.439	0.260	0.261	0.242	0.333	0.409
Avg.	0.396	0.273	0.303	0.391	0.403	0.269	0.226	0.247	0.362	0.418	0.272	0.240	0.217	0.306	0.335
GARCH	0.430	0.296	0.325	0.419	0.452	0.285	0.241	0.263	0.394	0.441	0.300	0.252	0.232	0.340	0.370
MS-GARCH-TVI	0.468	0.338	0.370	0.452	0.519	0.303	0.270	0.298	0.429	0.488	0.286	0.246	0.233	0.348	0.405
MS-GARCH-TVC	0.461	0.318	0.335	0.414	0.437	0.290	0.245	0.263	0.386	0.432	0.295	0.239	0.218	0.324	0.350
RealGARCH	0.237	**0.182**	**0.239**	0.380	0.409	0.256	0.208	0.229	0.358	0.408	0.331	0.261	0.229	**0.287**	**0.289**
HEAVY	0.272	0.223	0.326	0.591	0.759	0.262	0.228	0.273	0.498	0.593	0.339	0.305	0.313	0.535	0.642
HAR	0.234	0.189	0.254	**0.359**	0.385	0.243	0.212	0.238	0.374	0.430	0.340	0.257	0.230	0.371	0.470
HAR (lev.)	**0.226**	0.187	0.258	0.362	0.387	**0.232**	0.211	0.240	0.378	0.429	0.286	0.245	0.227	0.373	0.470
No-change	0.418	0.821	1.143	2.213	2.310	0.304	0.297	0.336	0.532	0.715	0.382	0.320	0.314	0.481	0.555

Notes: See Table 1.7 but for a confidence level of 95% instead of 90%.

2 Volatility forecasting for low-volatility investing

Abstract

Low-volatility investing is typically implemented by sorting stocks based on simple volatility proxies; for example, the empirical standard deviation of last year's daily returns. In contrast, we understand identifying next-month's ranking of volatilities as a forecasting problem aimed at the ex-post optimal sorting. We show that time series models based on intraday data outperform simple risk measures in anticipating the cross-sectional ranking of S&P 500 constituents in real time. The corresponding portfolios are more similar to the ex-ante infeasible optimal portfolio in multiple dimensions. However, even though some of the best models have higher returns than the benchmark, this holds only before transaction costs are taken into account.

2.1 Introduction

In the financial industry, low-risk strategies have become increasingly popular during recent years. Examples for those strategies are: betting against beta (Frazzini and Pedersen, 2014), low-volatility portfolios (Blitz and van Vliet, 2007), minimum variance portfolios (Clarke, de Silva, and Thorley, 2006), and volatility-managed portfolios (Moreira and Muir, 2017).

In this paper, we focus on the implementation of low-volatility portfolios. In financial practice, stocks are usually sorted according to some simple metric of a stock's total or idiosyncratic volatility. One example is the empirical standard deviation of monthly or daily returns over a certain period (the previous year, the previous 6-months, the previous month).[1] The corresponding low-volatility portfolio simply consists of, say, the 20% stocks with the lowest volatility. The portfolio is re-balanced on a monthly basis.

Clearly, from an ex-ante perspective it is not clear which proxy for stock volatility is best suited for stock selection. Therefore, we think of targeting the *optimal* low-volatility portfolio as

[1] Bali, Engle, and Murray (2016) provide an overview of the various metrics that are commonly used.

a forecasting problem. We first introduce the *oracle* low-volatility portfolio which we define as the portfolio that an investor would choose with hindsight. Following the literature on estimating volatility from high-frequency intraday return data, we measure the monthly volatility ex-post by realized variances (Andersen et al., 2003). Using data for all stocks in the S&P 500 and the period 2002–2018, we document the low-volatility anomaly from an ex-post perspective: low-volatility stocks have higher returns than high-volatility stocks.

We then investigate the question whether state-of-the-art volatility models are useful for anticipating the correct composition of the oracle portfolio in real time. That is, in each month we estimate various volatility models and use them to forecast the next month's volatility of each stock in the S&P 500. We then form low-volatility portfolios based on the sorting of stocks according to the forecasted volatilities.

During recent years there has been substantial progress in the development of volatility models. We use those recent models but also more established approaches. First, we use simple RiskMetrics models and various generalized autoregressive conditional heteroskedasticity (GARCH)-type models. In those models the conditional variance is treated as a latent process and daily (or monthly) returns are used for estimating volatilities. Second, we use heterogeneous autoregression (HAR)- and mixed-frequency data sampling (MIDAS)-type models. Here, the realized variances are modeled directly as a function of past realized variances. In addition, we consider forecast combinations; that is, we combine the forecasts from various volatility models according to measures of past forecast performance. We refer to those forecast combinations as "loss-based forecasts." We also use the measures that are commonly used for the volatility sorting of stocks as forecasting "models." For example, we consider the rolling window sample variance of daily returns based on the previous twelve months as the forecast for next month's volatility. We refer to those models (forecasts) as benchmark models (forecasts). We then compare the forecast performance of the volatility models with the forecast performance of those benchmark models.

For the evaluation of the forecast performance we take two alternative perspectives. The first one is common in the financial econometrics literature (e.g., Ghysels et al., 2019): For each stock we evaluate the forecast performance of each model and check which model performs best and how the volatility models compare with the benchmark models. Unsurprisingly, the volatility forecasts of state-of-the-art volatility models outperform the simple benchmark model when measuring forecast accuracy by standard loss functions. For example, for 33% of the stocks the best performing model (according to the squared error loss) is a HAR-type specification that also includes a variance forecast for the S&P 500. In general, the HAR models dominate GARCH-type models and the benchmark models are dominated by essentially all other models. However, identifying the "optimal" volatility model for each stock is only possible ex-post and not in real-time because of potential time-variation in model performance and the small sample period.

Alternatively, in each month we use a specific volatility model to forecast the volatilities of all stocks. Based on the cross-sectional forecast performance of each model, we select the optimal model on a period-by-period basis. This is our second perspective which is feasible in real time. Now, we find that the Realized GARCH is the best model in 21% of months (according to the squared error loss). Again, we find that GARCH- and HAR-type models dominate the benchmark models but the differences in performance are now somewhat weaker. The loss-based forecasts lead to further improvements in forecast performance.

Next, we investigate whether the model-based volatility forecasts allow us to construct low-volatility portfolios that are "closer" to the oracle portfolio than the portfolios that are based on the benchmark forecasts (henceforth benchmark portfolios). In that respect, it is important to note that it is not necessary to perfectly forecast each stocks' volatility in order to perfectly mimic the oracle portfolio. For example, if a model generates volatility forecasts which overestimate the volatility of each stock by 10%, the implied ordering of the stocks will still be fully correct. In addition, the empirical evidence in previous studies suggests that the relation between risk and return is rather flat for low- and medium-volatility stocks and then decreasing for high-volatility stocks (Blitz, van Vliet, and Baltussen, 2019). Hence, misclassifying stocks may not be that costly as long as we avoid to include high-volatility stocks in the portfolio. Our results suggest that portfolios which employ loss-based forecasts (henceforth loss-based portfolios) mimic the true oracle portfolio more closely than the benchmark portfolios. We measure "closeness" by the "oracle overlap;" that is, the time series average of the share of stocks that a particular low-volatility portfolio has in common with the oracle portfolio. At maximum we reach an oracle overlap of 67% which is more than 2.5 percentage points above the oracle overlap of the best benchmark portfolio. In that sense, the low-volatility portfolios that are based on state-of-the-art volatility models clearly improve upon the benchmark low-volatility portfolios.

However, when we compare the performance in terms of returns there are no significant differences between the model/loss-based portfolios and the *best* benchmark portfolio. There are two explanations. First, as mentioned before, certain misclassifications are not costly as long as severe classification errors are avoided. The best benchmark model which uses returns over the previous year appears to satisfy this criterion. In contrast, the benchmark model which is based on returns over the previous month only, has a relatively low oracle overlap (58%) and generates larger classification errors. Second, although volatilities are quite persistent, the oracle portfolio has a relatively high turnover (71%). As a consequence, the model/loss-based portfolios that achieve a high oracle overlap also generate a high turnover and, therefore, high transaction costs. Hence, the high oracle overlap comes at the cost of high transaction costs. As mentioned before, the best benchmark portfolio has a lower oracle overlap but also much lower turnover (only 16%) and, as a result, lower transaction costs. After trading costs, there are no significant differences in returns.

The rest of the article is structured as follows. In Section 2.2 we review the previous literature and present empirical evidence for the low-volatility anomaly. Section 2.3 presents the volatility models and Section 2.4 the data. We then evaluate the forecast performance of the volatility models in Section 2.5. A comparison of the various low-volatility portfolios is provided in Section 2.6. Finally, Section 2.7 concludes.

2.2 The low-volatility anomaly

2.2.1 Related literature

Since the 1970s, numerous empirical studies have shown that the risk-return relationship is either flat or even negative which is in contrast to the prediction of the CAPM. The anomaly holds irrespectively whether risk is defined to be beta (Black, Jensen, and Scholes, 1972; Haugen and Heins, 1972, 1975), total volatility (Haugen and Heins, 1972, 1975), or idiosyncratic volatility (Ang et al., 2006, 2009). This is due to the fact that on stock level total volatility is highly correlated with idiosyncratic volatility and high-beta stocks are typically high-volatility stocks (Baker, Bradley, and Wurgler, 2011; Blitz, van Vliet, and Baltussen, 2019).

Both rational and behavioral explanations have been proposed. One rational explanation is that investors face leverage constraints (Black, 1972); for example, regarding short-selling. Frazzini and Pedersen (2014) propose a model that incorporates such leverage constraints. Another rational explanation by Blitz and van Vliet (2007) argues that portfolio managers are typically subject to relative performance objectives which might render low-volatility stocks unattractive. A behavioral explanation is the possible preference of some investors for lottery-like payoffs examined by Barberis and Huang (2008) and Bali, Cakici, and Whitelaw (2011). Asness et al. (2020) find evidence that support both the leverage and the lottery hypothesis.

In contrast to the studies above, we examine the low-volatility anomaly from a forecasting perspective by employing time series models that are widely documented to perform better than trailing volatility. Ghysels, Santa-Clara, and Valkanov (2005) derive variance forecasts for the market based on mixed-data-sampling to provide evidence for a positive risk-return relationship. In a similar manner, Fu (2009) uses the exponential generalized autoregressive conditional heteroskedasticity model by Nelson (1991) to forecast idiosyncratic volatilities which he finds to be positively correlated with returns—contradicting Ang et al. (2006, 2009). The fact that total volatility predicts returns is also exploitable by machine-learning techniques as shown by Gu, Kelly, and Xiu (2020). In this regard, Ghysels, Santa-Clara, and Valkanov (2005), Fu (2009), and Gu, Kelly, and Xiu (2020) demonstrate the usefulness of time series models for portfolio sorting but their analyses are restricted to using daily return data.

The literature on intraday data for variance-based portfolio sorting follows the simple trailing volatility approach. The study by Boudt, Nguyen, and Peeters (2015) may be considered to be closest to ours. Like us, they use a S&P 500 real-time constituents data set to overcome the survivorship bias in De Pooter, Martens, and Van Dijk (2008) and Hautsch, Kyj, and Malec (2015). In their analysis, they come to the conclusion that there is no (statistically significant) benefit in returns from using intraday data but portfolio returns are less volatile. In contrast to our study, they do not use volatility models and have a short sample from 2007–2012. However, already Haugen and Heins (1975) note that high-volatility stocks are primarily outperformed by low-volatility stocks at longer investment periods which they attribute to superior performance during bear markets. Liu (2009) concludes that at a monthly investment horizon there is no benefit from intraday data if an investor has access to at least 12 months of daily data. Similarly, Amaya et al. (2015) find no significant predictive power of lagged realized variances on weekly stock returns.

Another branch of the literature examines volatility timing for aggregated portfolio returns (Moreira and Muir, 2017, 2019; Cederburg et al., 2020).

2.2.2 A new perspective on the anomaly

In this section, we take a new perspective on the low-volatility anomaly by creating and evaluating the performance of an ex-ante infeasible “oracle portfolio.” The usual approach in the literature on the low-volatility anomaly is as follows: At the end of each month m, all stocks are ranked according to a proxy of their volatility. Volatility is often measured as the square-root of the sum of squared daily returns over the previous month, the previous six months or the previous year (Bali, Engle, and Murray, 2016).[2] Based on the ranking for month m, equally weighted quintile portfolios for month $m+1$ are constructed. Then, according to the low-volatility anomaly the portfolio of stocks in the first quintile has higher average and risk-adjusted returns than the portfolio of stocks in the fifth quintile (e.g., Blitz and van Vliet, 2007).

We now take an ex-post perspective by asking the following question: What would have been the “true” quintile portfolios; that is, the portfolios that are formed based on the ex-post volatilities? Because stock volatilities are latent, even ex-post the correct ranking of stocks is not absolutely certain. We rely on the literature on estimating stock volatility from high-frequency intraday data and, hence, base the ex-post oracle portfolios on realized variances: At the end of each month $m + 1$, we compute the monthly realized volatility of each stock as the square-root of the sum of daily realized variances based on 5-minute intraday data and squared overnight returns.[3] We then consider the ex-ante infeasible quintile portfolios that are formed at the end of month m according to the realized volatility from the end of month $m + 1$. Although these quintile portfolios cannot be constructed in real-time, they tell us how an investor would have behaved with hindsight. Figure 2.1 shows the performance of the quintile portfolios during the 2002 to 2018 period. The first quintile portfolio clearly outperforms all other quintile portfolios. The weakest performance can be observed for the fifth quintile portfolio. Thus, the preliminary results for our oracle portfolio confirm the low-volatility anomaly from an ex-post perspective. In the following, we will refer to the first quintile portfolio as “the” oracle portfolio.[4]

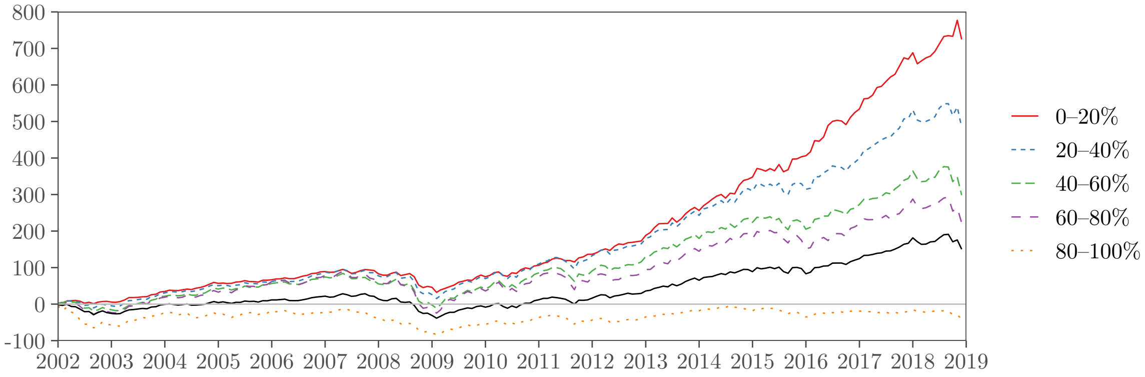
In Table 2.1, we compare the performance of the infeasible oracle portfolio with the performance of feasible low-volatility portfolios based on the volatility of the previous year (12m-RV^d), the previous six months (6m-RV^d) and the previous month (1m-RV^d). We will refer to those three portfolios as (feasible) benchmark portfolios. Table 2.1 shows that the infeasible oracle portfolio has a higher return, a lower volatility and, hence, a higher Sharpe ratio than the three

[2] Because total volatility and idiosyncratic volatility are typically highly correlated, Bali, Engle, and Murray (2016) argue that portfolio sortings on one or the other measure of volatility usually lead to the same results. Using a measure based on data from the last month was suggested in Ang et al. (2006, 2009).

[3] For details see Section 2.4.

[4] As an alternative oracle portfolio, we considered a portfolio that is based on a monthly volatility measure which uses squared daily returns only. However, the oracle portfolio based on intraday realized variances clearly outperforms the portfolio based on squared daily returns in terms of average excess returns and Sharpe ratio. It also has considerably lower turnover.

Figure 2.1: Discrete returns of oracle volatility portfolios.



Notes: Monthly discrete excess returns of the quintile oracle portfolios based on all S&P 500 constituents; 2002:M1 to 2018:M12. As a benchmark, excess returns of the S&P 500 Total Return Index are depicted in black.

benchmark portfolios. Note that the oracle as well as the benchmark portfolios clearly beat the market portfolio in all three dimensions. While the oracle portfolio achieves a risk reduction of 33%, the benchmark portfolios reach a risk reduction of almost 30%.

Table 2.1: Summary of oracle and benchmark portfolio.

	Avg	Std	SR	ARVol	OO	SO	TO
Oracle	12.92	9.44	1.37	20.45	—	65.29	70.87
12m-RV^d	8.39	10.03	0.84	23.44	64.32	93.58	15.62
6m-RV^d	8.33	10.07	0.83	23.40	64.93	88.71	25.09
1m-RV^d	8.35	10.56	0.79	24.18	58.48	52.45	95.94
S&P 500 TR	6.42	14.12	0.45	—	—	—	—

Notes: In Panel A, we report arithmetic means of discrete excess returns (Avg), their standard deviation, and the corresponding Sharpe ratio (SR). ARVol is the square-root of the time-averaged "average realized variance" which is defined to be the cross-sectional average RV inside the corresponding low-volatility portfolio, $ARVol = \sqrt{\frac{1}{MN}\sum_{m=1}^{M}\sum_{i=1}^{N} RV_{i,m}}$. Annualized scale. Oracle overlap is the average share of ex-post oracle stocks that are included in the benchmark portfolio. Self-overlap (SO) is the average share of stocks staying in the corresponding low-volatility portfolio. For the definition of turnover (TO) see Subsection 2.6.3. OO, TN, and SO are reported in percentages. The portfolios are based on the S&P 500 constituents in between 2002:M1–2018:M12.

At first sight, there seem to be no major differences in the performance of the three benchmark portfolios. However, differences become apparent when considering additional characteristics of the portfolios. First, we compute the average realized volatility (ARVol) of each portfolio. That is, in each month m we compute the cross-sectional average of the realized variance of the stocks in the portfolio and then average over time. ARVol is the square-root of this quantity. By construction, the oracle portfolio has the lowest ARVol. While the ARVol figures for 6m-RV^d

and 12m-RV^d are similar, the ARVol figure for 1m-RV^d is the highest which suggests that the 1m-RV^d portfolio has the severest classification errors. This is confirmed when computing the "oracle overlap" (OO): In each month m we count how many of the stocks that are included in the benchmark portfolios are also part of the oracle portfolio. We average the corresponding share over time. On average, only 58.44% of the stocks in the 1m-RV^d portfolio are also part of the oracle portfolio. This number increases to almost 65% for the 6m-RV^d and 12m-RV^d portfolios. Next, we compute the self-overlap (SO) for each portfolio. We define the SO as the average number of stocks that stay in the low-volatility portfolio from one month to the next. Here, the differences between the benchmark portfolios become much more pronounced. While the SO of the 1m-RV^d portfolio is only 52% the SO of the 12m-RV^d portfolio is almost 94%. This is due to the fact that the ranking of the stocks' volatility based on the previous month is much more volatile than the ranking based on the previous year. From a practical perspective, this makes a huge difference because the corresponding turnover (TO) of the two portfolios is 96.03% and 15.62% respectively. This implies that after transaction costs the 12m-RV^d portfolio clearly dominates the 1m-RV^d portfolio (see Section 2.6.4). Hence, in the following, we will refer to the 12m-RV^d portfolio as the "benchmark portfolio." Although the TO of the oracle portfolio is comparably high, we will show that even after (reasonable) transaction costs it generates higher returns than any of the benchmark portfolios.

Obviously, an investor would be interested in replicating the oracle portfolio as closely as possible. We denote the realized variance of stock i, $i = 1, \ldots, n$, in month $m + 1$ by $RV_{i,m+1}$. The oracle portfolio is based on the ascending ordering of the monthly realized variances of all n stocks: $RV_{1,m+1} \leq RV_{2,m+1} \leq \ldots \leq RV_{n,m+1}$. Hence, we can think of the task of replicating the oracle portfolio as a forecasting problem. We forecast the realized variances of the n stocks based on information up to the end of month m and form a portfolio based on the ranking that is implied by the forecasted variances $\widehat{RV}_{i,m+1}$, $i = 1, \ldots, n$. We will address the forecasting problem in three steps:

1. We first estimate various volatility models for each stock and evaluate the forecast performance of each model. This allows us to answer the following questions: Do state-of-the-art volatility models provide better forecasts of the cross-sectional stock volatility than the simple benchmark models? Is there a single volatility model (or a few volatility models) that outperform(s) the others? Because the benchmark models are not designed to accu-

rately forecast volatility but rather to "identify" stocks that qualify for the low-volatility portfolio, we expect that the answer to the first question will be "yes." As most of the literature on volatility forecasting focuses on daily forecasts, the one-month horizon that is needed in our setting will shed some new light on the potential advantages of models that directly model the realized variances over models that treat the conditional variance as latent when forecasting volatility over longer horizons.

2. We will evaluate whether the forecasts from the volatility models do translate into a "more accurate" ranking of stock volatilities than the forecasts from the benchmark models. We will measure the accuracy by the oracle overlap. That is, we evaluate whether the decision to include a stock in the low-volatility portfolio is correct. Note that the oracle overlap can be high, even if the ranking that is implied by the volatility forecasts is far from perfect. However, a perfect ranking would imply a 100% oracle overlap.

3. Do the portfolios with the highest oracle overlap generate the highest returns? We will see that the answer to this question crucially depends on portfolio turnover and transaction costs.

2.3 Models

We consider a wide range of models which represent the state of the art in volatility modeling. The models can be broadly classified as either RiskMetrics, GARCH, HAR or MIDAS. While in the GARCH and RiskMetrics approach volatility is treated as a latent variable, the HAR and MIDAS specifications model realized variances directly. In the following, we briefly introduce the various model specifications. A more detailed description of the different models can be found in Appendix 2.8.1.

RiskMetrics (RM): We use four variants of the RiskMetrics model. Two variants employ monthly realized variances based on squared daily returns while the other two employ weighted averages of squared daily returns directly. The RiskMetrics models use either six or twelve months of past return data. Note that the RiskMetrics models can be considered as restricted GARCH models with fixed ARCH/GARCH parameters and a constant equal to zero.

GARCH: Besides the simple GJR GARCH of Glosten, Jagannathan, and Runkle (1993), we employ a "Panel GARCH" model which uses variance targeting for each stock and restricts the

ARCH/GARCH coefficients to be the same across stocks. We also use the Factor GARCH model of Engle, Ng, and Rothschild (1990) and combine it with the GARCH-MIDAS of Engle, Ghysels, and Sohn (2013). As explanatory variables in the long-term component, we use the VIX, housing starts and the term spread. Those variables have been shown to be powerful predictors of longer term volatility (Conrad and Loch, 2015; Conrad and Kleen, 2020). Correspondingly, these models are denoted as Factor GARCH-VIX, Factor GARCH-ΔHous, and Factor GARCH-TS. We also consider the Realized GARCH as suggested in Hansen, Huang, and Shek (2012) and two types of multiplicative error (MEM) models (Engle and Gallo, 2006).

HAR: We consider the original HAR specification as suggested by Corsi (2009) as well as seven extensions. In the original HAR model the realized variance is a linear function of the lagged daily, weekly, and monthly realized variances. Among the extensions are specifications that model the realized variance of stock i as depending on stock i's lagged realized variances but also on a HAR-type forecast for the S&P 500, or the VIX index. We also use the "Panel HAR" model of Bollerslev et al. (2018).

MIDAS: This type of volatility model has been proposed in Ghysels, Santa-Clara, and Valkanov (2004, 2005, 2006). The realized variance is modeled as a weighted average of lagged daily realized variances. The weights are parsimoniously parameterized via a flexible parametric weighting scheme. The HAR model of Corsi (2009) is nested when imposing certain constraints on the weights.

We estimate all models on a rolling window of four years with a minimum number of 600 observations.[5] Forecasts are computed for month $m = 1, \ldots, M$.

Ghysels et al. (2019) study the performance of iterated versus direct multi-step ahead forecasting for GARCH, HAR and MIDAS models. Following their recommendations, we directly forecast the average 22-day realized variance for all HAR-type models. Similarly, we construct direct forecasts for the MIDAS models. The GARCH and MEM models are estimated using daily data and then iterated volatility forecasts are computed.

[5] The only exceptions are three variants of Factor GARCH-MIDAS models which employ housing starts or term spread data beginning in 1987 and the VIX and S&P 500 returns beginning in 1990 in order to identify the long-term component.

2.4 Data

Monthly portfolio returns are calculated from monthly CRSP total returns and the real-time constituents list for the S&P 500 is downloaded from Compustat. We adjust for CRSP delisting returns such that we have a survivorship bias free data set (Shumway, 1997; Bali, Engle, and Murray, 2016).

Our data provider of one-minute intraday data for individual stocks is QuantQuote.[6] One-minute intraday data for the S&P 500 is downloaded from Tick Data. Daily values for the VIX and monthly returns for the S&P 500 Total Return Index are obtained from the Cboe website.[7] We estimate all time series models and evaluate our forecasts on the daily/intraday data set. The first date of observations is January 02, 1998 and the last date is December 31, 2018. For the intraday realized variance estimates, we include prices during market hours from 9:30 to 16:00 and calculate 5-minute log-returns. The first 5-minute return of each day is an open-close return and all others are close-close ones. We use 5-minute returns for two reasons: First, because this frequency is commonly used, it makes our analysis comparable. Second, it has been shown to be a fairly robust choice as a trade-off between using high-frequency data and obstructing micro-structure noise related estimation errors (Liu, Patton, and Sheppard, 2015). To further strengthen our proxy, we average across subsampled 5-minute realized variances starting 9:30, 9:31, 9:32, 9:33, and 9:34. In order to have a measure on the daily scale, we add squared overnight returns to the intraday realized variance. At day t and for stock i we will denote this combined measure by $RV_{i,t}$. The monthly realized variance, $RV_{i,m}$, of stock i is the sum of $RV_{i,t}$ over all days t in month m. Alternatively, squared daily (close-close) returns are often used as a simple but less accurate measure of volatility. We will denote it by $RV^d_{i,t}$.

Discrete excess market returns $R_{mkt,t}$ and the corresponding risk-free rates $R_{rf,t}$ are obtained from Kenneth R. French's data library.[8] For further factor analyses, we use the Fama-French(-Carhart) four- and five-factor portfolio returns; that is, daily average returns of SMB (Small Minus Big), HML (High Minus Low), MOM (Momentum), RMW (Robust Minus Weak) and CMA (Conservative Minus Aggresive) portfolios (Fama and French, 1993; Carhart, 1997; Fama and French, 2015). These are also obtained from Kenneth R. French's data library website.

[6] Similarly, Bollerslev, Li, and Zhao (2019) and Bollerslev, Patton, and Quaedvlieg (2020) merge CRSP with NYSE TAQ data.

[7] http://www.cboe.com/products/vix-index-volatility/vix-options-and-futures/vix-index/vix-historical-data and http://www.cboe.com/micro/buywrite/monthendpricehistory.xls

[8] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/

SMB_t is the return on a diversified portfolio of small stocks minus the return on a diversified portfolio of big stocks, HML_t is the difference between the returns on diversified portfolios of high and low book-to-market ratio stocks The two additional factors in the five-factor model can be understood as measures of profitability and investment. Hence, RMW_t is the calculated as the difference between returns on diversified portfolios of stocks with robust and weak profitability, and CMA_t is calculated as the difference between returns on diversified portfolios of low and high investment stocks, which Fama and French call conservative and aggressive. Because of collinearity, Fama and French (2015) refrain from including the momentum effect in their five-factor model and we follow their approach.

Last, real-time housing starts data is downloaded from ALFRED[9] and term-spread data from the New York Federal Reserve website.[10]

Due to the data restrictions from our rolling estimation scheme detailed in Section 2.3, we include on average 480 S&P 500 constituents in our portfolio selection. In total we have 97,940 monthly stock returns in the investment period from 2002–2018.

2.5 Forecast evaluation and model selection

In a first step, we evaluate the volatility forecasts from the different models. In the following subsection, we introduce four loss functions and then provide empirical results from an ex-post and a real-time perspective.

2.5.1 Loss functions

Following Patton (2011), we evaluate the volatility forecasts using robust loss functions. Suppose we are interested in evaluating the conditional variance forecast $\widehat{RV}_{m+1|m}$ against the true but unobservable conditional variance σ^2_{m+1} using the loss function $L(\sigma^2_{m+1}, \widehat{RV}_{m+1|m})$.[11]

Then, the loss function is called robust if the expected loss ranking of two competing forecasts is preserved when replacing σ^2_{m+1} by a conditionally unbiased proxy. In the empirical application, we use the monthly realized variances RV_{m+1} as proxies for the unobservable σ^2_{m+1}. We will

[9] https://alfred.stlouisfed.org/series?seid=HOUST

[10] https://www.newyorkfed.org/research/capital_markets/ycfaq.html#/

[11] In this subsection, for simplicity in the notation we drop the index i.

employ two popular loss functions which are robust: the squared error (SE) loss,

$$L(\sigma^2_{m+1}, \widehat{RV}_{m+1|m}) = (\sigma^2_{m+1} - \widehat{RV}_{m+1|m})^2,$$

and the QLIKE loss,

$$L(\sigma^2_{m+1}, \widehat{RV}_{m+1|m}) = \sigma^2_{m+1}/\widehat{RV}_{m+1|m} - \log(\sigma^2_{m+1}/\widehat{RV}_{m+1|m}) - 1.$$

As a third loss function, we consider the elementary loss (EL). For a pre-specified threshold θ, the EL assigns a penalty if and only if $\widehat{RV}_{m+1|m}$ is below/above θ while σ^2_{m+1} is above/below θ:

$$L(\sigma^2_{m+1}, \widehat{RV}_{m+1|m}) = \begin{cases} |\sigma^2_{m+1} - \theta| & \text{if } \widehat{RV}_{m+1|m} \leq \theta < \sigma^2_{m+1} \\ |\sigma^2_{m+1} - \theta| & \text{if } \sigma^2_{m+1} \leq \theta < \widehat{RV}_{m+1|m} \\ 0 & \text{else.} \end{cases}$$

More generally, all loss functions that belong to the so-called class of Bregman loss functions satisfy the conditions for robustness (Patton, 2011). As Ehm et al. (2016) show that any Bregman loss function can be expressed as an integral of elementary losses, we know that the EL is also robust. In the case of low-volatility investing, a natural choice for θ is the 20%-quantile ($\theta^{(20)}$) of the cross-sectional distribution of stock volatilities. Thus, we only penalize forecast errors with respect to the targeted threshold of $\theta^{(20)}$ in each month and denote the losses by EL 20.

Finally, we rely on the cross-sectional Mincer-Zarnowitz R^2 as a measure of forecast accuracy (Mincer and Zarnowitz, 1969). This is the R^2 from a cross-sectional regression of RV_{m+1} on $\widehat{RV}_{m+1|m}$ (henceforth MZ R^2). If the MZ R^2 is equal to one, then we have perfectly forecasted the ranking of the volatilities.[12] The MZ R^2 has also been shown to be robust (Hansen and Lunde, 2006). We report the loss from the MZ R^2 as $1/R^2$ so that lower means better as it is the case for the SE, QLIKE, and EL.

The differences between the four evaluation criteria can be summarized as follows: While the SE is a symmetric loss function, the QLIKE is asymmetric and penalizes underestimation more heavily than overestimation. The QLIKE is less affected by extreme observations and

[12] Theoretically, the MZ R^2 would also be equal to one if we perfectly forecasted the reverse ranking but this is of no concern in our application.

requires weaker moment conditions when performing Diebold-Mariano-type tests (Patton, 2006). Empirically, the SE and QLIKE are based on the average forecast losses across all observations whereas the EL assigns and averages (non-zero) losses only for those stocks which were falsely included into/excluded from the low-volatility portfolio. Contrary to the previous three loss functions, we can think of the MZ R^2 as directly evaluating the entire forecast ranking.

2.5.2 Ex-post perspective

First, we evaluate the forecast performance of the various volatility models from an ex-post perspective. For each stock i, we consider the out-of-sample volatility forecasts $\widehat{RV}^{j}_{i,m|m-1}$, $m = 1, \ldots, M$, stemming from model j. For each loss function and with hindsight, we can measure the average loss of model j for stock i across time as

$$L_i^j = \frac{1}{M} \sum_{m=1}^{M} L^j(RV_{i,m}, \widehat{RV}^{j}_{i,m|m-1}).$$

We denote the stock specific loss of the benchmark forecast (12m-RV^d) by L_i^B. As a measure for the forecast accuracy of a particular model j relative to the benchmark, we consider the following statistic

$$LR_i^j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{L_i^j/L_i^B<1},$$

where $\mathbf{1}_{L_i^j/L_i^B<1}$ is an indicator function which equals one if $L_i^j/L_i^B < 1$ and zero else. Hence, LR_i^j reports the share of stocks for which model j outperforms the benchmark. Table 2.2 shows LR_i^j for the four loss functions. Independently of the loss function, almost all models beat the benchmark for more than 50% of the stocks. In particular, we find that HAR-type models perform very well relative to the benchmark. For example, the SE loss of the HAR-SPX-LR model is lower than the loss of the benchmark (12m-RV^d) for 93% of the stocks. Among the GARCH-type models the Realized GARCH does best according to the SE. To the contrary, the simple MEM appears to be outperformed by the benchmark. Interestingly, the RiskMetrics model based on monthly returns (RM monthly) and twelve months performs very well but for the EL 20.

In order to compare the various models not only to the benchmark but also with each other, we also report the share of stocks for which a particular model j performed best (denoted by $Rk_i^j \leq 1$) or was among the best four models (denoted by $Rk_i^j \leq 4$) as measured by L_i^j. Our

Table 2.2: Ex-post comparison of model performance.

	SE			QLIKE			EL 20			MZ R^2		
Model	LR_i^j	$Rk_i^j \leq 1$	$Rk_i^j \leq 4$	LR_i^j	$Rk_i^j \leq 1$	$Rk_i^j \leq 4$	LR_i^j	$Rk_i^j \leq 1$	$Rk_i^j \leq 4$	LR_i^j	$Rk_i^j \leq 1$	$Rk_i^j \leq 4$
12m-RV^d	—	0.00	0.03	—	0.00	0.02	—	0.00	0.04	—	0.01	0.02
6m-RV^d	0.74	0.01	0.03	0.71	0.00	0.02	0.58	0.00	0.04	0.87	0.01	0.03
1m-RV^d	0.48	0.00	0.01	0.30	0.00	0.01	0.47	0.01	0.04	0.80	0.03	0.05
RM monthly, 12 months	0.93	0.01	0.03	0.93	0.01	0.02	0.34	0.00	0.03	0.90	0.00	0.02
RM monthly, 6 months	0.76	0.01	0.03	0.71	0.00	0.03	0.59	0.00	0.05	0.87	0.00	0.03
RM daily, 12 months	0.75	0.00	0.03	0.67	0.01	0.07	0.68	0.01	0.07	0.89	0.00	0.03
RM daily, 6 months	0.74	0.00	0.03	0.65	0.00	0.06	0.67	0.01	0.07	0.89	0.01	0.04
GJR-GARCH	0.75	0.01	0.05	0.79	0.01	0.08	0.72	0.03	0.09	0.79	0.02	0.06
Panel GJR-GARCH	0.71	0.03	0.07	0.72	0.03	0.08	0.67	0.02	0.09	0.83	0.03	0.07
Factor GARCH	0.85	0.01	0.07	0.85	0.02	0.13	0.75	0.02	0.07	0.86	0.01	0.05
Factor GARCH-VIX	0.84	0.02	0.08	0.64	0.01	0.03	0.47	0.01	0.05	0.87	0.01	0.06
Factor GARCH-ΔHous	0.84	0.01	0.07	0.66	0.01	0.03	0.46	0.01	0.05	0.86	0.00	0.04
Factor GARCH-TS	0.82	0.00	0.06	0.64	0.00	0.03	0.42	0.00	0.04	0.84	0.00	0.03
Realized GARCH	0.91	0.09	0.18	0.85	0.09	0.17	0.61	0.05	0.11	0.91	0.08	0.15
MEM	0.36	0.00	0.01	0.01	0.00	0.00	0.11	0.01	0.03	0.60	0.01	0.03
Panel MEM	0.65	0.01	0.05	0.04	0.00	0.01	0.21	0.01	0.03	0.80	0.01	0.05
HAR	0.86	0.01	0.12	0.90	0.01	0.17	0.81	0.03	0.18	0.89	0.01	0.09
HAR-SPX	0.91	0.17	0.64	0.92	0.06	0.41	0.81	0.05	0.30	0.93	0.16	0.64
HAR-LR	0.90	0.03	0.20	0.94	0.10	0.53	0.84	0.07	0.35	0.94	0.04	0.22
HAR-SPX-LR	0.93	0.33	0.76	0.94	0.27	0.70	0.82	0.11	0.46	0.96	0.39	0.79
Panel HAR	0.74	0.00	0.01	0.70	0.00	0.01	0.72	0.01	0.11	0.81	0.00	0.02
Panel HAR-LR	0.83	0.02	0.05	0.87	0.00	0.05	0.79	0.03	0.15	0.88	0.01	0.05
HAR-VIX	0.91	0.10	0.68	0.91	0.16	0.60	0.81	0.06	0.35	0.94	0.06	0.67
HAR-VIX-LR	0.92	0.12	0.69	0.91	0.18	0.63	0.81	0.09	0.42	0.96	0.10	0.73
MIDAS	0.72	0.00	0.01	0.85	0.00	0.02	0.77	0.01	0.08	0.83	0.00	0.01
Panel MIDAS	0.78	0.00	0.03	0.86	0.01	0.08	0.76	0.04	0.16	0.86	0.01	0.02

Notes: LR_i^j reports the share of losses L_i^j to be smaller than L_i^B; this is, the proportion of stocks for which the loss of the respective model j is smaller than the one of the 12m-RV^d benchmark model. $Rk_i^j \leq 1$ and $Rk_i^j \leq 4$ report the proportion of the model being the best or among the four best-performing models as measured by L_i^j. The evaluation is based on the cross-section of S&P 500 constituents in between 2002:M1–2018:M12.

previous findings are confirmed: According to the SE loss, the HAR-SPX-LR model has the lowest loss for 33% of the stocks. Other models that perform well are the HAR-SPX, the HAR-VIX and the HAR-VIX-LR. Again, the best GARCH-type model is the realized GARCH. When considering the top-4 models and according to the SE, the HAR-SPX-LR model is included in this set for 76% of stocks. Interestingly, the three benchmark models are almost never among the top-4. Note that the ranking of models is relatively robust across loss functions. In summary, HAR-type models clearly dominate when forecast performance is evaluated for each stock separately from an ex-post perspective; that is, based on the time series of out-of-sample forecast errors for each stock. Our finding is largely in line with Ghysels et al. (2019).

2.5.3 Real-time perspective

The model and stock specific losses L_i^j are available ex-post only. Hence, we cannot use them in the real-time portfolio selection process.[13] Instead, we will rely on cross-sectional forecast losses. That is, for each loss function and model j, we define the cross-sectional average loss in

[13] Of course, it is possible to compute the losses L_i^j for rolling/expanding windows of out-of-sample forecasts and select models based on this. However, given the monthly frequency of the forecasts and our sample period, model selection will be difficult due to the small sample that is used to compute a rolling/expanding window version of L_i^j.

month m as:[14]

$$L_m^j = \frac{1}{n}\sum_{i=1}^{n} L^j(RV_{i,m}, \widehat{RV}_{i,m|m-1}^j).$$

We denote the loss of the benchmark model by L_m^B. The losses L_m^j and L_m^B can be used in real-time for the selection of models. Ex-post, we can also compute the statistic

$$LR_m^j = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{L_m^j/L_m^B<1}.$$

where $\mathbf{1}_{L_m^j/L_m^B<1}$ equals one if $L_m^j/L_m^B < 1$ and zero else. Hence, LR_m^j reports the share of months during which model j outperforms the benchmark. As Panel A of Table 2.3 shows, most model-based forecasts still beat the benchmark forecast from a cross-sectional perspective. However, forecast improvements are less impressive with LR_m^j often being slightly above 50%. While the HAR-SPX-LR still performs very well, the Realized GARCH has a slightly higher LR_m^j statistic. In general, the HAR-type models are now less dominant. In addition, we now report for how many months a specific model j is ranked top (denoted by $Rk_m^j \leq 1$) or among the top-4 models (denoted by $Rk_m^j \leq 4$) in terms of L_m^j. Independent of the loss function, the Realized GARCH is most often the best model. The Realized GARCH and the HAR-SPX-LR are most often among the top-4 models. However, we observe that many models are among the top-4 in more than 10% of months. That is, from a cross-sectional perspective we do not find that one specific model dominates all others. Also note that the 12m-RV^d and 6m-RV^d benchmark models are among the top-4 in a non-negligible number of months.

Thus, the real-time forecast evaluation suggests either that the differences between the various models are less pronounced from a cross-sectional perspective or that the forecast performance of the different models varies over time. The latter could be the case if one model is particularly suited for a high-volatility environment while another one performs best in a low-volatility environment (Conrad and Kleen, 2020). In the following, we consider forecast combinations as a means to safeguard against such time-varying model performance. The idea to combine the forecasts from different models to achieve diversification gains was popularized by Bates and Granger (1969). For further discussions see, for example, Timmermann (2006).

We follow the approach described in Caldeira et al. (2017) for combining the forecasts of the various models. First, for each model j we determine the cross-sectional forecast performance

[14] For simplicity in the notation, we assume that the number of stocks, n, in the cross-section is fixed.

Table 2.3: Real-time comparison of model performance.

		SE			QLIKE			EL 20			MZ R^2		
		LR_m^j	$Rk_m^j \leq 1$	$Rk_m^j \leq 4$	LR_m^j	$Rk_m^j \leq 1$	$Rk_m^j \leq 4$	LR_m^j	$Rk_m^j \leq 1$	$Rk_m^j \leq 4$	LR_m^j	$Rk_m^j \leq 1$	$Rk_m^j \leq 4$
Panel A: Model-based forecasts													
12m-RV^d		—	0.01	0.13	—	0.03	0.17	—	0.03	0.20	—	0.03	0.22
6m-RV^d		0.51	0.03	0.17	0.48	0.05	0.18	0.50	0.03	0.20	0.46	0.02	0.17
1m-RV^d		0.31	0.01	0.02	0.24	0.00	0.00	0.26	0.00	0.03	0.20	0.00	0.01
RM monthly, 12 months		0.78	0.04	0.15	0.72	0.05	0.15	0.53	0.04	0.18	0.79	0.07	0.25
RM monthly, 6 months		0.51	0.03	0.17	0.49	0.03	0.20	0.52	0.09	0.20	0.49	0.03	0.19
RM daily, 12 months		0.54	0.01	0.11	0.49	0.00	0.18	0.50	0.03	0.18	0.43	0.02	0.08
RM daily, 6 months		0.54	0.00	0.08	0.47	0.01	0.14	0.47	0.03	0.19	0.43	0.00	0.07
GJR-GARCH		0.54	0.01	0.04	0.54	0.00	0.06	0.62	0.03	0.16	0.43	0.01	0.07
Panel GJR-GARCH		0.44	0.00	0.05	0.48	0.02	0.10	0.55	0.02	0.14	0.38	0.00	0.02
Factor GARCH		0.51	0.00	0.02	0.50	0.01	0.05	0.58	0.03	0.16	0.38	0.00	0.03
Factor GARCH-VIX		0.60	0.02	0.17	0.44	0.01	0.09	0.40	0.02	0.11	0.37	0.00	0.05
Factor GARCH-ΔHous		0.59	0.01	0.12	0.46	0.01	0.10	0.43	0.02	0.11	0.38	0.00	0.04
Factor GARCH-TS		0.59	0.01	0.19	0.48	0.01	0.09	0.41	0.01	0.10	0.38	0.00	0.03
Realized GARCH		0.77	0.21	0.44	0.70	0.20	0.35	0.64	0.16	0.33	0.72	0.25	0.46
MEM		0.37	0.03	0.08	0.11	0.00	0.03	0.04	0.00	0.01	0.18	0.00	0.01
Panel MEM		0.39	0.06	0.10	0.29	0.08	0.13	0.23	0.06	0.10	0.27	0.03	0.08
HAR		0.60	0.01	0.10	0.60	0.04	0.17	0.57	0.01	0.14	0.56	0.02	0.11
HAR-SPX		0.66	0.04	0.23	0.61	0.04	0.24	0.55	0.03	0.16	0.57	0.05	0.24
HAR-LR		0.67	0.04	0.25	0.64	0.09	0.32	0.58	0.06	0.24	0.64	0.05	0.34
HAR-SPX-LR		0.73	0.11	0.40	0.64	0.08	0.33	0.56	0.03	0.22	0.67	0.11	0.45
Panel HAR		0.45	0.02	0.07	0.41	0.02	0.05	0.46	0.01	0.07	0.44	0.00	0.06
Panel HAR-LR		0.51	0.01	0.09	0.47	0.01	0.10	0.51	0.03	0.13	0.56	0.02	0.16
HAR-VIX		0.70	0.04	0.33	0.63	0.02	0.25	0.51	0.02	0.18	0.63	0.05	0.27
HAR-VIX-LR		0.72	0.17	0.41	0.63	0.12	0.34	0.51	0.07	0.21	0.66	0.14	0.40
MIDAS		0.48	0.00	0.00	0.50	0.00	0.03	0.56	0.00	0.09	0.46	0.00	0.03
Panel MIDAS		0.52	0.02	0.09	0.55	0.03	0.14	0.56	0.07	0.16	0.52	0.05	0.15
Panel B: Combined forecasts													
$\eta = 0$		0.78	—	—	0.81	—	—	0.72	—	—	0.75	—	—
$\eta = 1/2$	SE	0.78	—	—	0.80	—	—	0.71	—	—	0.74	—-	—
	QLIKE	0.78	—	—	0.79	—	—	0.71	—	—	0.75	—-	—
	EL 20	0.78	—	—	0.79	—	—	0.72	—	—	0.74	—-	—
	MZ R^2	0.79	—	—	0.80	—	—	0.71	—	—	0.75	—-	—
$\eta = 1$	SE	0.78	—	—	0.79	—	—	0.71	—	—	0.74	—	—
	QLIKE	0.78	—	—	0.80	—	—	0.72	—	—	0.75	—	—
	EL 20	0.77	—	—	0.79	—	—	0.71	—	—	0.72	—	—
	MZ R^2	0.79	—	—	0.79	—	—	0.71	—	—	0.75	—	—
$\eta = 4$	SE	0.79	—	—	0.79	—	—	0.70	—	—	0.75	—	—
	QLIKE	0.79	—	—	0.78	—	—	0.68	—	—	0.74	—	—
	EL 20	0.75	—	—	0.74	—	—	0.62	—	—	0.69	—	—
	MZ R^2	0.80	—	—	0.80	—	—	0.70	—	—	0.77	—	—
$\eta = \infty$	SE	0.71	—	—	0.63	—	—	0.53	—	—	0.64	—	—
	QLIKE	0.76	—	—	0.68	—	—	0.57	—	—	0.67	—	—
	EL 20	0.72	—	—	0.65	—	—	0.58	—	—	0.63	—	—
	MZ R^2	0.79	—	—	0.72	—	—	0.66	—	—	0.72	—	—

Notes: LR_m^j reports the proportion of months in which the cross-sectional loss L_m^j of model j is lower than the one of the 12m-RV^d benchmark forecast. $Rk_m^j \leq 1$ and $Rk_m^j \leq 4$ report the proportion of the model being the best or among the four best-performing models as measured by L_m^j. The evaluation is based on the cross-section of S&P 500 constituents in between 2002:M1–2018:M12.

at month m as

$$\bar{L}_m^j = \frac{1}{m} \sum_{k=0}^{m-1} \delta^k L_{m-k}^j, \tag{2.1}$$

with $\delta \in (0, 1]$. When δ approaches zero, we exclusively rely on the loss ratio in month m. In the other extreme, when $\delta = 1$, the forecast performance is measured by the simple average of the loss ratios over the previous m months. For $0 < \delta < 1$ all loss ratios are taken into account but the weights are declining from the most recent to the most distant observation in time. Throughout the main analysis we will set $\delta = 0.98$.[15]

[15] In Appendix 2.8.2 we report alternative returns for $\delta \in \{0, 0.6, 0.8, 0.9, 0.94, 0.99, 0.999, 1\}$. Returns are slightly

The combined forecast for the volatility of stock i, $i = 1, \ldots, n$, in period $m+1$ is given by

$$\widehat{RV}^{cf}_{i,m+1|m} = \sum_{j=1}^{J} \lambda_{j,m} \widehat{RV}^{j}_{i,m+1|m}, \tag{2.2}$$

where the weights are given by

$$\lambda_{j,m} = \frac{(\bar{L}^j_m)^{-\eta}}{\sum_{j=1}^{J} (\bar{L}^j_m)^{-\eta}}. \tag{2.3}$$

with $\eta \geq 0$. For $\eta = 0$, we attach equal weights, $\lambda_{j,m} = 1/J$, to each model. For $\eta = \infty$ a weight of one is attached to the model for which the loss in Equation (2.1) is the lowest and all other models receive a weight of zero. When $\eta = 1$, the weights are inverse proportional to the loss of the respective model. Note that $\eta = 1/2$ in combination with the SE means that the weights are chosen according to the root mean squared error.

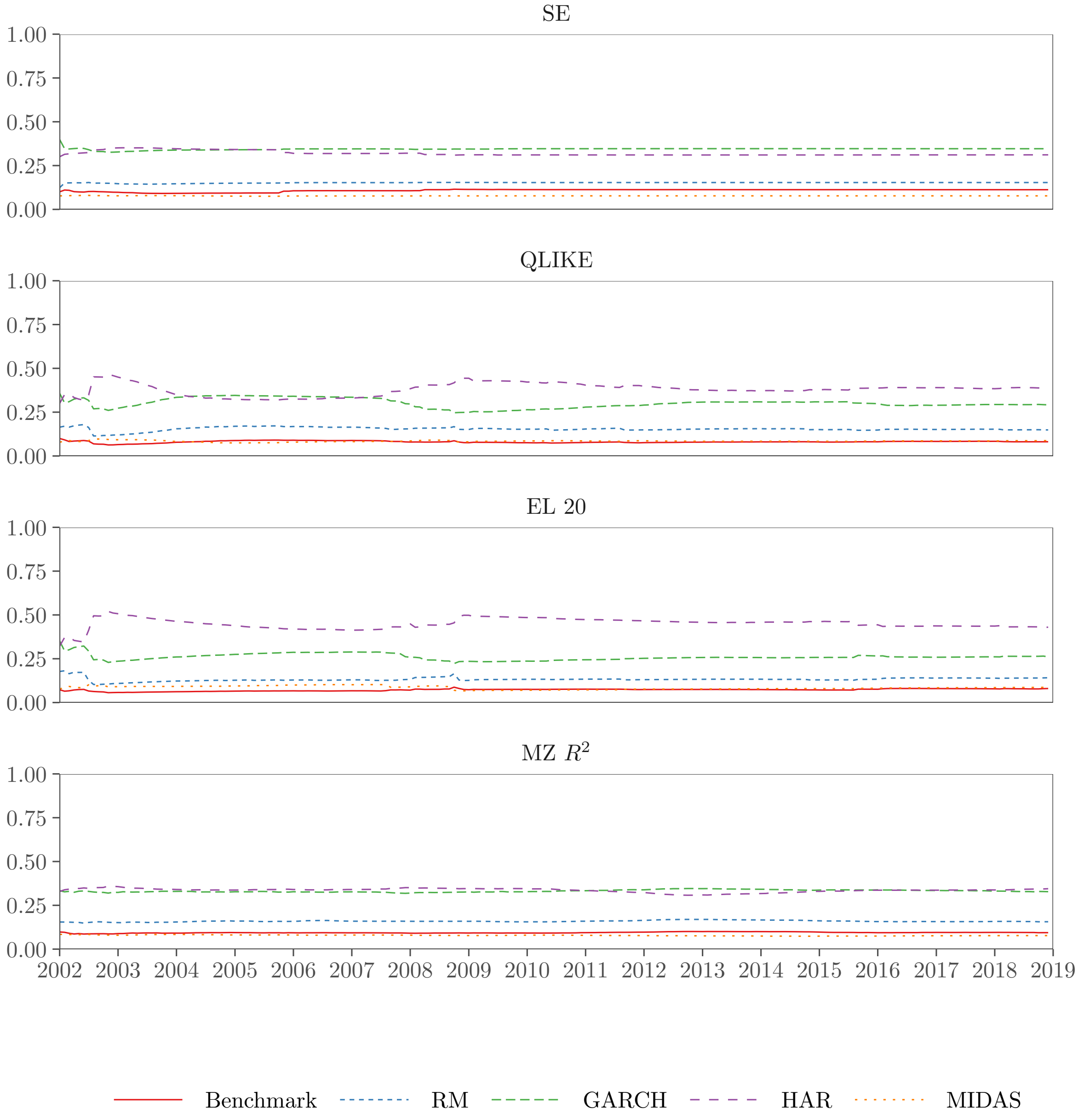
In Figure 2.2, we plot the time series of weights that are attached to each model class when choosing $\eta = 1$ and $\delta = 0.98$. For example, the green line shows the cumulative weights that are attached to all GARCH-type models. The figure shows an interesting dichotomy: the SE and the MZ R^2 attach roughly the same weight (around 30%) to the HAR- and GARCH-type models. The RiskMetrics models receive slightly more weight than the benchmark models. In contrast, the QLIKE and, in particular, EL 20 assign substantially more weight to the HAR models than to the GARCH models. In addition, while the plots for the SE and the MZ R^2 suggest that the relative forecast performance is constant over time, the plots for the QLIKE and EL 20 imply that there is some time-variation in forecast performance.

Figure 2.3 illustrates how the weights change for $\eta = 4$. According to Equation (2.3), we now give stronger weights to those models with superior forecast performance. As a result, the disparity among the weights increases. In particular, this is the case for the QLIKE and the EL 20. Now, the QLIKE attaches an aggregated weight of up to 75% to the HAR models which dominate the other model classes since the financial crisis in 2008. An even more distinctive weighting can been seen for the EL 20. Beginning in 2003, the HAR models have a joint weight of at least 60% and sometimes even more than 90%.

Panel B of Table 2.3 presents the forecast performance of the combined forecasts. There is a remarkable finding which holds independently of the choice of the loss function and the choice of η: For almost all combined forecasts the statistic LR^{cf}_m is considerably higher than for the

higher for $\delta = 1$ but we choose $\delta = 0.98$ as it seems to be a reasonable choice an investor could have made ex-ante.

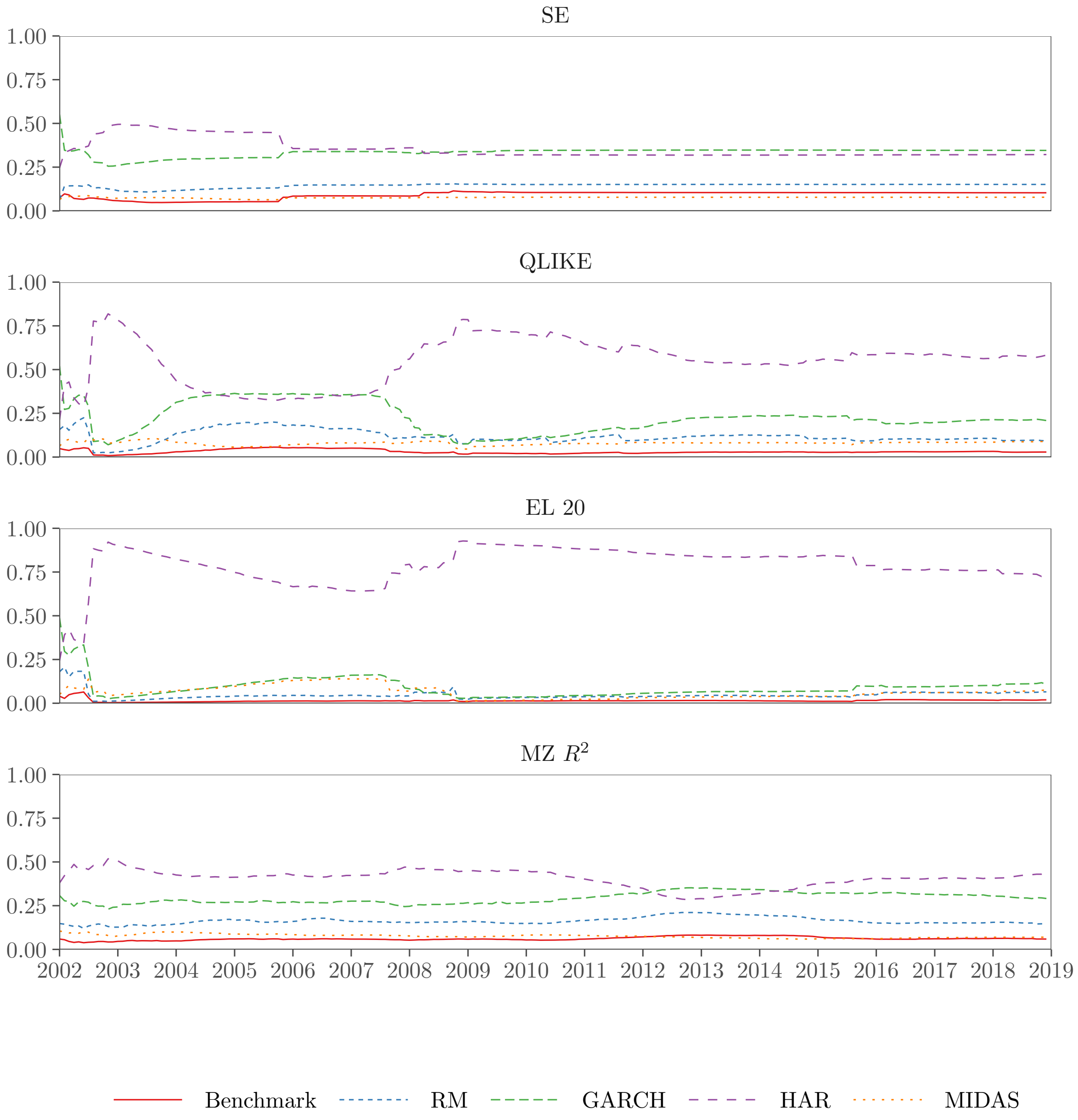
Figure 2.2: Weight of model class per month for $\eta = 1, \delta = 0.98$.



Notes: Aggregated weights in the combined forecast per model class, see Equations (2.2) and (2.3). The constituents of each class are listed in 2.3. The class "Benchmark" includes 12m-RV^d, 6m-RV^d, and 1m-RV^d.

individual model-based forecasts. That is, the dominance of the combined forecasts over the benchmark forecast is much stronger than for the individual models. For $\eta = \infty$ the advantage is less striking because in each month now all weight is attached to one specific model which reduces the potential diversification gains. Thus, from a pure forecasting perspective it clearly pays off to consider the combined forecast. The finding that the loss function which is used to combine the individual model-based forecasts does not seem to matter much is highly interesting.

Figure 2.3: Weight of model class per month for $\eta = 4, \delta = 0.98$.



Notes: See notes of Figure 2.2 but for $\eta = 4, \delta = 0.98$.

The theoretical arguments that can be made in favor or against certain loss functions appear not being relevant in our setting. Even a simple average ($\eta = 0$) of the forecasts appears to do the job.

2.6 Comparison of low-volatility portfolios

2.6.1 Portfolio construction

We illustrate the construction of the low-volatility portfolios for volatility forecasts based on model j. Assume that the volatility forecasts $\widehat{RV}^j_{i,m|m-1}$ for the n stocks in month m are already in ascending order; that is, $\widehat{RV}^j_{1,m|m-1} \leq \widehat{RV}^j_{2,m|m-1} \leq \ldots \leq \widehat{RV}^j_{n,m|m-1}$. Based on this ordering of the forecasts, the 20% stocks with the lowest volatility are included in the portfolio for month m. Those stocks receive equal weights (i.e., a weight of one over the number of stocks in the portfolio). All remaining stocks receive a weight of zero. We denote the individual weights by $w^j_{m|m-1}$.

When constructing low-volatility portfolios the decision whether a particular stock is included in the portfolio or not solely depends on the ascending ordering of the forecasted volatilities of all stocks. Hence, for correctly mimicking the oracle portfolio it is not necessary to perfectly forecast volatility. All that matters is an accurate ranking of the stocks' volatility. However, a perfect forecast leads to an accurate ranking. Hence, we conjecture that volatility models which provide more accurate forecasts should also deliver a more accurate ranking of the volatilities.

2.6.2 ARVol and oracle overlap

For each low-volatility portfolio the column denoted ARVol in Table 2.4 shows the time series mean of the average cross-sectional volatility in each month. Recall that the oracle portfolio has an ARVol of 20.45%. Among the model-based portfolios, the HAR-based portfolios achieve the lowest ARVol. The best-performing model is the HAR with an ARVol of 22.9%. With the exception of the $\eta = \infty$ case, all loss-based portfolios (i.e. the portfolios which are based on the combined forecasts) achieve lower ARVol's than the HAR model. The ARVol of the 12m-RV^d benchmark is 23.44%.

The column denoted OO in Table 2.4 shows the oracle overlap of the low-volatility portfolios that are either based on the volatility forecasts of a single model (Panel A) or the combined forecasts (Panel B). The benchmark portfolio has an average overlap of 64.32%. That is, in each month m and based on the volatility forecasts of the benchmark model for month $m+1$ we decide whether or not to include a specific stock in the low volatility portfolio. The ex-post comparison with the oracle portfolio in month $m+1$ shows that on average the decision was correct for 64.32% of the stocks. Panel A shows that the model-based portfolios generally do not

Table 2.4: Portfolio characteristics.

		ARVol	OO	SO	TO
Panel A: Model-based portfolios					
12m-RV^d		23.44	64.32	93.58	15.62
6m-RV^d		23.40	64.93	88.71	25.09
1m-RV^d		24.18	58.48	52.45	95.94
RM monthly, 12 months		23.43	64.58	93.43	15.89
RM monthly, 6 months		23.36	65.05	88.70	25.08
RM daily, 12 months		23.36	64.72	79.03	43.87
RM daily, 6 months		23.37	64.57	78.74	44.44
GJR-GARCH		23.71	62.00	79.55	42.73
Panel GJR-GARCH		23.74	61.97	77.38	46.92
Factor GARCH		23.73	61.40	86.24	29.92
Factor GARCH-VIX		23.88	59.86	84.04	34.21
Factor GARCH-ΔHous		23.82	60.02	83.65	34.98
Factor GARCH-TS		23.83	59.92	83.86	34.57
Realized GARCH		23.31	65.95	80.83	40.50
MEM		24.92	54.31	83.24	36.79
Panel MEM		23.60	60.91	81.93	40.01
HAR		22.90	64.10	82.79	36.69
HAR-SPX		22.93	63.35	84.50	33.47
HAR-LR		22.92	63.97	81.56	39.12
HAR-SPX-LR		22.98	63.27	81.52	39.31
Panel HAR		23.39	63.46	88.80	25.00
Panel HAR-LR		23.28	64.99	91.00	20.71
HAR-VIX		22.98	63.37	83.95	34.51
HAR-VIX-LR		23.12	62.61	81.10	40.11
MIDAS		23.53	63.49	83.03	36.26
Panel MIDAS		23.22	65.79	85.58	31.22
Panel B: Loss-based portfolios					
$\eta = 0$		22.70	66.94	87.02	28.26
$\eta = 1/2$	SE	22.71	66.87	87.06	28.17
	QLIKE	22.69	66.88	87.24	27.83
	EL 20	22.67	66.96	87.30	27.70
	MZ R^2	22.70	66.90	87.35	27.62
$\eta = 1$	SE	22.71	66.86	87.04	28.20
	QLIKE	22.67	66.90	87.37	27.57
	EL 20	22.56	66.95	87.46	27.42
	MZ R^2	22.69	66.93	87.55	27.23
$\eta = 4$	SE	22.70	66.89	87.09	28.12
	QLIKE	22.53	66.80	87.42	27.54
	EL 20	22.53	65.98	86.72	28.98
	MZ R^2	22.56	67.05	88.09	26.19
$\eta = \infty$	SE	22.95	63.51	81.20	39.80
	QLIKE	23.02	63.28	81.36	39.59
	EL 20	22.98	63.29	81.73	38.86
	MZ R^2	22.88	65.14	79.69	42.76

Notes: Summary measures of the model-based and loss-based portfolios are reported. ARVol is the annualized square-root of the time-averaged cross-sectional realized variance inside each portfolio, see notes in Table 2.1. Oracle overlap (OO), self-overlap (SO), and turnover (TO) are reported in percentages, see Subsection 2.6.2 and 2.6.3. The evaluation is based on the cross-section of S&P 500 constituents in between 2002:M1–2018:M12.

improve upon the benchmark portfolio. In contrast, with the exception of $\eta = \infty$, all loss-based portfolios lead to improvements. Their ARVol is close to 67%. The best loss-based portfolio uses the MZ R^2 in combination with $\eta = 4$ and has an oracle overlap of 67.05%. Hence, the improved forecast performance of the combined forecasts leads to improvements in the oracle overlap of more than 2.5 percentage points relativ to the benchmark portfolio. Again, the simple average ($\eta = 0$) of all model-based forecasts does surprisingly well.

2.6.3 Self-overlap, portfolio turnover and transaction costs

As we are interested in measuring the actual performance of our low-volatility portfolios, we need to take into account the accruing transaction costs when implementing the strategy. As an intermediate step, we report the self-overlap of each portfolio in column SO in Table 2.4. The self-overlap of the benchmark 12m-RV^d portfolio is 93.58%. In sharp contrast, most model-based portfolios achieve only self-overlap of around 80%. This is due to the fact that the model-based forecasts are typically not as persistent as the benchmark forecasts. This drawback is partially addressed by the loss-based portfolios. As Panel B shows, the loss-based portfolios have a higher self-overlap of around 87%. One exception is the loss-based portfolio for $\eta = \infty$. In this case, in each month the best single model-based forecast achieves a weight of one and, hence, the forecasts are less persistent and the corresponding portfolio has lower self-overlap.

The previous findings suggest that the model-based forecasts and the loss-based forecasts should generate a higher portfolio turnover and thereby higher transaction costs than the benchmark model. We compute the turnover and the respective transaction costs following the recent literature on portfolio-allocation based on high-frequency-based measures of realized (co-)variation (Bandi, Russell, and Zhu, 2008; De Pooter, Martens, and Van Dijk, 2008; DeMiguel, Garlappi, and Uppal, 2009; Hautsch, Kyj, and Malec, 2015; Nolte and Xu, 2015). Recall that $w^j_{i,m|m-1}$ is either zero or one divided by the number of stocks in the portfolio. Before the next rebalancing at the end of period m, due to price movements, the weight of stock i changes to $w^j_{i,m|m-1}\frac{1+R_{i,m}/100}{1+(w^j_m)'R_m/100}$ where $w^j_m = (w^j_{1,m}, \dots, w^j_{n,m})'$ and $R_m = (R_{1,m}, \dots, R_{n,m})'$. Based on the volatility forecasts for month $m + 1$, the new desired weights are $w^j_{i,m+1|m}$. Hence, the turnover due to portfolio rebalancing at the end of month m is given by

$$TO^j_m = \sum_{i=1}^{n} \left| w^j_{i,m+1|m} - w^j_{i,m|m-1}\frac{1 + R_{i,m}/100}{1 + (w^j_m)'R_m/100} \right| .$$

The quantity TO_m can be interpreted as the proportion of wealth reallocated at the end of month m. In column TO, we report the average turnover for each portfolio. The turnover of the benchmark portfolio is 15.62% which means that per dollar invested the average transaction volume per month is 15.62 cents. Relying on model-based forecasts increases the turnover for most models to be in the 25%–40% range. The highest turnover is observed for the naive 1m-RV^d forecast. Except for the case of $\eta = \infty$, the loss-based portfolios have a turnover of roughly 27%.

In summary, most loss-based portfolios outperform the benchmark in terms of oracle overlap but not self-overlap and turnover. Hence, we expect that trading costs will hurt the model- and loss-based portfolio performance. This is what we investigate next.

2.6.4 Portfolio returns

Assuming transaction costs to be proportional to the portfolio turnover TO_m, we follow DeMiguel, Garlappi, and Uppal (2009) and compute monthly portfolio excess returns as

$$R^j_{p,m} = \frac{W^j_m}{W^j_{m-1}} - 1 - R_{rf,m},$$

where W^j_m is the wealth of the model/loss-based portfolio which can be obtained as

$$W^j_m = W^j_{m-1} \cdot (1 + w'_m R_m) \cdot (1 - c \cdot TO_m).$$

We assume that c is ranging from 0 to 25bps which is a realistic range of recent cost estimates for trading large US stocks (Novy-Marx and Velikov, 2016).

Table 2.5 shows the annualized returns of each portfolio for $c \in \{0, 15, 25\}$. When there are no transaction costs, the benchmark portfolio earns an annualized return of 8.39%. For the model and loss-based portfolio, we report the annualized return and, in brackets, the p-value of a t-test using Newey-West standard errors that checks whether there is a significant difference between the return of the respective model/loss-based portfolios and the benchmark. Although the returns of most of the model/loss-based portfolios are somewhat higher than the return of the benchmark, we do not find evidence for a significant difference besides for the HAR-SPX-LR and HAR-VIX-LR, two of the best-performing models in Subsection 2.5.3. Hence, even without transaction costs the superior performance of model (loss-based) volatility forecasts does not necessarily translate into higher returns. The same holds for the standard deviation of returns and the Sharpe ratios.

Once we take transaction costs into account the picture is clearly more in favor of the benchmark 12m-RV^d portfolio. For c equal to 15bps the return of the benchmark portfolio falls to 8.10% but only five HAR models generate returns higher than that. Because the turnover of the benchmark is much lower than the turnover of the model/loss-based portfolios, its returns are less affected.

An alternative strategy targeted at conservative investors are buy-and-hold portfolios. However, given that our asset-universe changes over time as companies enter or leave the S&P 500, we can only compare ourselves to strategies that invest in a passive index-tracking fund. For example, the average return of the S&P 500 Total Return Index, in which dividend-payments are included, is 6.42% annually with a Sharpe ratio of 0.45 during our investment period (see Table 2.1). Under the assumption of an expense ratio of around 0.1%, which is the current rate of the SPDR S&P 500 ETF Trust, we see that even after transaction costs of 25bps all our loss-based strategies generate returns more than 1 percentage point higher than a buy-and-hold strategy on the S&P 500. This achievement comes not at the cost of higher volatility as all loss-based portfolios have standard deviations close to 4 percentage points lower than the S&P 500 Total Return Index with a volatility of 14.12%. As a result of earning higher average returns while reducing volatility, the Sharpe ratios of the low-volatility portfolios are almost twice as large as the proposed buy-and-hold benchmark. The same holds for the 12m-RV^d portfolio.

2.6.5 Utility analysis

We follow Fleming, Kirby, and Ostdiek (2001, 2003) and evaluate the various portfolios in a utility-based framework. This allows us to judge whether the differences between the benchmark and the model/loss-based portfolios are of economic significance. Using a quadratic utility function with risk-aversion parameter γ, the monthly utility generated by a portfolio based on model j is given by

$$U_\gamma(R^j_{p,m}) = (1 + R^j_{p,m}/100) - \frac{\gamma}{2(1+\gamma)}\left(1 + R^j_{p,m}/100\right)^2 .$$

We are now interested in comparing this utility with the utility from the oracle portfolio. Denote the return of the oracle portfolio by $R^o_{p,m}$. We can compute the maximum fee Δ^j_γ that an investor would be willing to pay in order to switch from portfolio j to the oracle portfolio by solving

$$\sum_{m=1}^{M} U_\gamma(R_{p,m}^j) = \sum_{m=1}^{M} U_\gamma(R_{p,m}^o - \Delta_\gamma^j). \tag{2.4}$$

The smaller Δ_γ^j the closer the model j based portfolio mimics the utility of the oracle portfolio. We report the fee Δ_γ^j in Table 2.5 in annualized percentage points for $\gamma = 1$ and $\gamma = 10$. Again, the model/loss-based portfolios outperform the benchmark portfolio only before transaction costs in utility terms. We observe the lowest fees for the HAR-SPX-LR based portfolio.

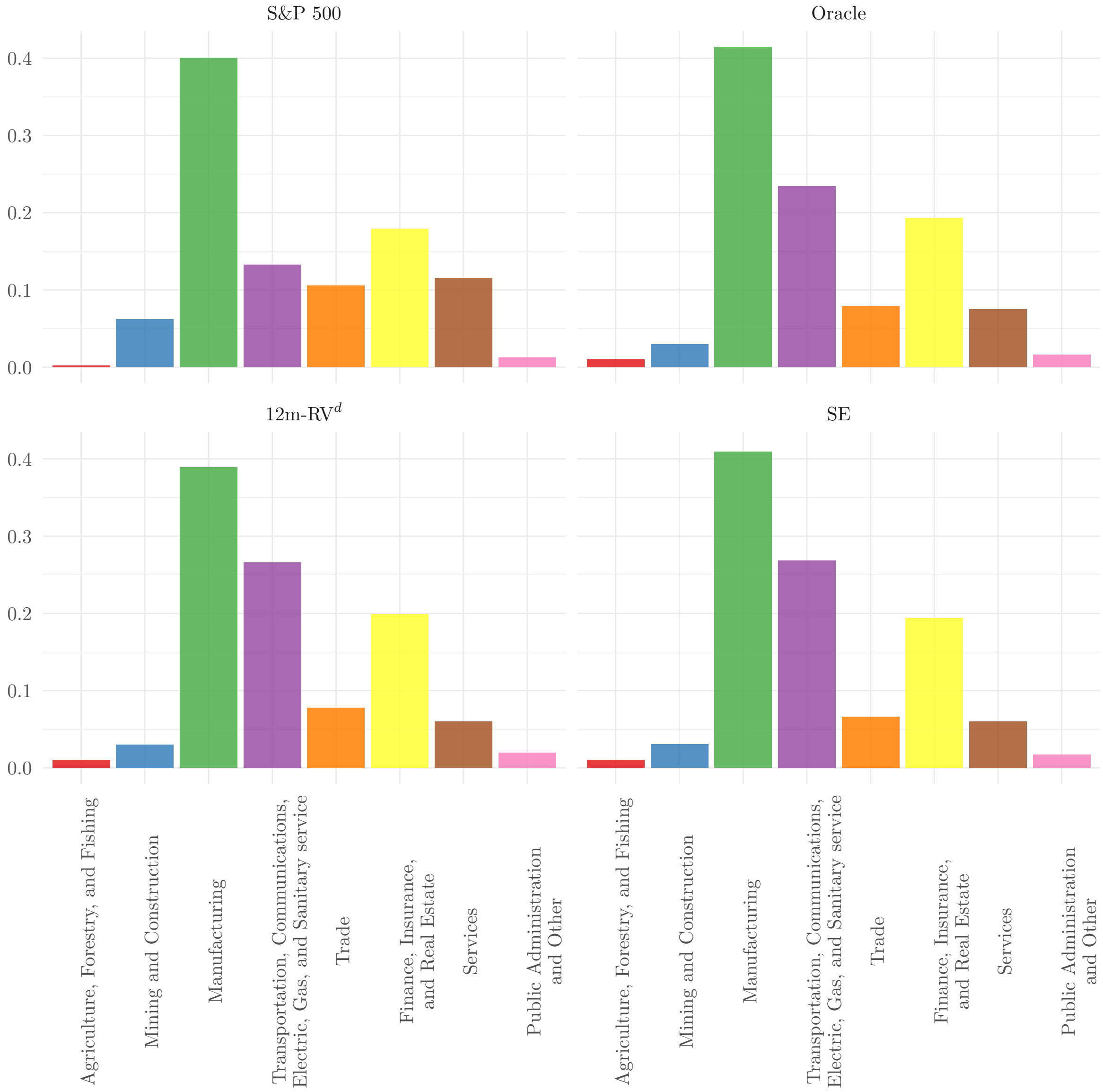
2.6.6 Sector concentration

We now examine whether our low-volatility investing strategies may generate high exposure to narrow classes of industries. In Figure 2.4 we depict histograms of the average sector concentration by primary SIC codes. We report numbers for the entire S&P 500 cross-section (upper left) and the low-volatility oracle, the 12m-RV^d benchmark, and the SE-based portfolio with $\eta = \infty, \delta = 0.98$. For brevity, the latter is considered to be representative for our model-based strategies. In Figure 2.4, we see that 40% of the S&P 500 constituents are classified as "Manufacturing."[16] The second largest industry is the "Finance, Insurance, and Real Estate" sector (18%), followed by "Transportation, Communications, Electric, Gas and Sanitary service" (13%), "Services" (12%), "Trade" (11%), and "Mining and Construction" (6%). Less than 1.5% of S&P 500 constituents are classified as "Public Administration" and "Forestry and Farming." We use realtime SIC codes from the CRSP files in order to allow companies to be reassigned to a new sector. One example is S&P Global Inc., formerly McGraw-Hill Companies, for which industry classification changes from "Printing and Publishing," which is part of the "Manufacturing"-sector, to "Security and Commodity Brokers" in "Finance, Insurance, and Real Estate" after the acquisition of financial service providers like SNL financial in April 2015 and divestures like the sale of McGraw-Hill Education in 2013.

The other three histograms of our low-volatility portfolios show that the higher returns do not come at the cost of overexposure to one particular sector. The share of the large "Manufacturing"-sector in our low-volatility portfolios is the same as in the aggregate S&P 500 which can also be observed for the "Finance, Insurance, and Real Estate"-sector. The largest absolute difference to the composition of the S&P 500 can be seen in the weight on the

[16] SIC code 2 and 3, see https://www.osha.gov/pls/imis/sic_manual.html.

Figure 2.4: Sector concentration of S&P 500 vs. oracle and low-volatility portfolios.



Notes: Sector concentration by realtime Standard Industrial Classification (SIC). We report the time-average proportion in the S&P 500 and the infeasible oracle portfolio along the corresponding numbers for two exemplary ex-ante feasible portfolios; the benchmark 12m-RV^d and the SE-based portfolio with $\eta = \infty, \delta = 0.98$. Industries are classified by the first number of the SIC code as follows: "Agriculture, Forestry and Fishing" (0), "Mining and Construction" (1), "Manufacturing" (2 and 3), "Transportation, Communications, Electric, Gas and Sanitary service" (4), "Trade" (5), "Finance, Insurance, and Real Estate" (6), "Services" (7 and 8), "Public Administration and Other" (9). The evaluation period is 2002:M1–2018:M12.

"Transportation, Communications, Electric, Gas and Sanitary service"-sector which increases by around 10 percentage points for all low-volatility portfolios. Noteworthy is also the decrease in weight for the sector "Mining and Construction." The weight for this sector is a good example of the difference between our low-volatility portfolios and minimum-variance portfolios: The mining companies are high-risk stocks but they exhibit only low correlation with other stocks (Blitz, van Vliet, and Baltussen, 2019). Hence, in a minimum-variance portfolio it may be sensible to include such high-risk but low-correlation stocks in order to minimize the overall portfolio risk. However, given the histograms we can conclude that both the benchmark and our loss-based strategies do not generate excess returns on the downside of excessive sector exposure.

2.6.7 Factor analysis

In Table 2.6, we evaluate the trading strategies of the 12m-RV^d benchmark and our loss-based portfolios for the exemplary case of $\eta = \infty$ and $\delta = 0.98$ by means of Fama-French regressions. First, we observe a statistically significant CAPM-excess return for both the benchmark portfolio and the loss-based portfolios which is around half in size relative to the total portfolio return. In the Fama-French-Carhart (FFC) four-factor model (Fama and French, 1993; Carhart, 1997), we observe a significant negative coefficient for the SMB portfolio returns which is in line with the observation that low-volatility is correlated with high market capitalization. Similarly, momentum also helps to partially explain the superior performance of the low-volatility strategies. However, the average FFC-excess returns of our strategies are only slightly below the ones for the CAPM. Turning to the Fama-French five-factor model (Fama and French, 2015), we see that the excess returns are not as good captured by size but its exposure to highly profitable but conservative investment stocks. The FF five-factor model implies a reduction in monthly excess returns by around one-third. However, with values in range of 2.5–3.1% the annualized excess returns are still statistically significant with a p-value of at most 2%. Using daily returns of the entire CRSP cross-section but a longer evaluation period, Fama and French (2016) report similar results for total (and idiosyncratic) volatility portfolios.

2.7 Conclusion

We examine the effect of employing intraday data and corresponding volatility models on long-only low-volatility investments. The portfolio choice problem at hand is to identify the bottom quintile of stocks with the lowest volatility among S&P 500 constituents. In general, the anomaly is exploited by sorting based on last year's volatility which we employ as our benchmark. However, the benchmark is at odds with the financial econometrics literature that demonstrated repeatedly the usefulness of intraday data for volatility forecasting; in particular, for short-term forecasting.

First, we show that a large number of different time series models based on intraday data have superior forecasting performance at a monthly horizon in comparison to our benchmark in the years 2002–2018. Our set of models includes Riskmetrics, numerous GARCH- and HAR-type models, and MIDAS regressions. The best-performing model is a HAR model that includes a long-run and a market component. Interestingly, the overall dominance of the HAR-type models across stocks is more pronounced if the models are evaluated on a stock-by-stock basis instead of a monthly cross-sectional perspective. In general, forecast performance improves after combining model-based forecasts in real time. Our forecast evaluation is robust against using different loss functions.

Second, it is revealed that superior forecast performance translates into better assessment of the volatility ranking. This is measured both in terms of lower realized variances across stocks inside the low-volatility portfolios and a larger overlap with the infeasible oracle portfolio. Loss-based forecast combination is also beneficial in terms of similarity to the oracle portfolio. However, even though some of the best models have higher returns than the benchmark, they are typically not significantly higher and do not survive transaction costs due to higher turnover.

Table 2.5: Returns of low-volatility portfolios depending on transaction costs.

		0 bps					15 bps					25 bps				
		Avg	Std	SR	Δ_1	Δ_{10}	Avg	Std	SR	Δ_1	Δ_{10}	Avg	Std	SR	Δ_1	Δ_{10}
Oracle		12.92	9.44	1.37	—	—	11.63	9.41	1.24	—	—	10.77	9.40	1.15	—	—
Panel A: Model-based portfolios																
12m-RV^d		8.39	10.03	0.84	4.60	5.18	8.10	10.03	0.81	3.59	4.19	7.91	10.03	0.79	2.92	3.53
6m-RV^d		8.33 [0.87]	10.07	0.83	4.65	5.28	7.88 [0.47]	10.07	0.78	3.82	4.46	7.58 [0.28]	10.07	0.75	3.26	3.91
1m-RV^d		8.35 [0.96]	10.56	0.79	4.69	5.84	6.61 [0.05]	10.54	0.63	5.14	6.27	5.45 [0.00]	10.53	0.52	5.44	6.56
RM monthly, 12 months		8.38 [0.92]	9.99	0.84	4.60	5.15	8.09 [0.87]	9.99	0.81	3.60	4.16	7.90 [0.84]	10.00	0.79	2.93	3.51
RM monthly, 6 months		8.25 [0.68]	10.07	0.82	4.73	5.36	7.80 [0.35]	10.07	0.77	3.90	4.54	7.50 [0.20]	10.08	0.74	3.34	4.00
RM daily, 12 months		8.02 [0.43]	10.16	0.79	4.97	5.69	7.23 [0.05]	10.17	0.71	4.48	5.22	6.70 [0.01]	10.17	0.66	4.15	4.90
RM daily, 6 months		8.14 [0.59]	10.11	0.81	4.85	5.52	7.33 [0.09]	10.11	0.73	4.37	5.05	6.80 [0.01]	10.12	0.67	4.04	4.74
GJR-GARCH		8.38 [0.98]	10.29	0.81	4.63	5.49	7.60 [0.22]	10.30	0.74	4.12	5.00	7.09 [0.04]	10.31	0.69	3.77	4.67
Panel GJR-GARCH		8.05 [0.50]	10.26	0.78	4.96	5.78	7.20 [0.08]	10.26	0.70	4.52	5.35	6.63 [0.01]	10.26	0.65	4.23	5.07
Factor GARCH		8.60 [0.61]	10.32	0.83	4.41	5.31	8.06 [0.91]	10.33	0.78	3.67	4.57	7.70 [0.61]	10.33	0.75	3.17	4.08
Factor GARCH-VIX		8.42 [0.95]	10.45	0.81	4.61	5.64	7.80 [0.51]	10.45	0.75	3.94	4.98	7.38 [0.25]	10.45	0.71	3.49	4.53
Factor GARCH-ΔHous		8.62 [0.60]	10.46	0.82	4.41	5.45	7.99 [0.79]	10.46	0.76	3.75	4.80	7.56 [0.43]	10.47	0.72	3.31	4.37
Factor GARCH-TS		8.67 [0.54]	10.49	0.83	4.37	5.44	8.04 [0.89]	10.50	0.77	3.70	4.79	7.62 [0.53]	10.50	0.73	3.26	4.35
Realized GARCH		7.94 [0.23]	10.13	0.78	5.05	5.74	7.21 [0.02]	10.13	0.71	4.49	5.20	6.72 [0.00]	10.13	0.66	4.12	4.83
MEM		8.27 [0.87]	10.92	0.76	4.81	6.36	7.60 [0.49]	10.93	0.70	4.18	5.74	7.16 [0.30]	10.93	0.65	3.77	5.33
Panel MEM		8.60 [0.68]	10.26	0.84	4.40	5.22	7.88 [0.67]	10.25	0.77	3.84	4.67	7.40 [0.32]	10.25	0.72	3.46	4.29
HAR		8.44 [0.90]	10.36	0.81	4.58	5.51	7.78 [0.46]	10.38	0.75	3.95	4.91	7.34 [0.19]	10.39	0.71	3.53	4.51
HAR-SPX		8.72 [0.44]	10.29	0.85	4.29	5.15	8.11 [0.99]	10.30	0.79	3.61	4.49	7.71 [0.63]	10.31	0.75	3.15	4.05
HAR-LR		8.96 [0.25]	10.36	0.86	4.06	5.00	8.25 [0.77]	10.38	0.79	3.48	4.44	7.78 [0.79]	10.39	0.75	3.09	4.07
HAR-SPX-LR		9.14 [0.09]	10.35	0.88	3.87	4.80	8.43 [0.45]	10.36	0.81	3.29	4.24	7.96 [0.92]	10.37	0.77	2.91	3.87
Panel HAR		8.39 [0.99]	10.41	0.81	4.63	5.61	7.94 [0.72]	10.41	0.76	3.79	4.79	7.64 [0.54]	10.42	0.73	3.24	4.24
Panel HAR-LR		8.27 [0.79]	10.40	0.80	4.75	5.73	7.90 [0.63]	10.41	0.76	3.83	4.83	7.65 [0.53]	10.41	0.73	3.22	4.23
HAR-VIX		8.91 [0.18]	10.19	0.87	4.09	4.85	8.28 [0.65]	10.20	0.81	3.43	4.20	7.87 [0.91]	10.21	0.77	2.98	3.77
HAR-VIX-LR		9.13 [0.07]	10.35	0.88	3.88	4.80	8.41 [0.46]	10.35	0.81	3.32	4.25	7.92 [0.98]	10.36	0.76	2.94	3.89
MIDAS		8.15 [0.65]	10.46	0.78	4.88	5.92	7.49 [0.25]	10.48	0.71	4.25	5.31	7.06 [0.11]	10.49	0.67	3.82	4.91
Panel MIDAS		8.05 [0.45]	10.34	0.78	4.97	5.88	7.48 [0.17]	10.34	0.72	4.24	5.16	7.11 [0.07]	10.34	0.69	3.76	4.69
Panel B: Loss-based portfolios																
$\eta = 0$		8.43 [0.89]	10.15	0.83	4.56	5.28	7.92 [0.57]	10.15	0.78	3.79	4.51	7.58 [0.30]	10.16	0.75	3.27	4.01
$\eta = 1/2$	SE	8.33 [0.86]	10.16	0.82	4.67	5.39	7.82 [0.38]	10.16	0.77	3.89	4.62	7.48 [0.18]	10.17	0.74	3.37	4.11
	QLIKE	8.40 [0.96]	10.17	0.83	4.60	5.33	7.90 [0.54]	10.18	0.78	3.81	4.56	7.56 [0.30]	10.18	0.74	3.29	4.05
	EL 20	8.33 [0.87]	10.17	0.82	4.67	5.40	7.83 [0.42]	10.18	0.77	3.88	4.63	7.49 [0.22]	10.18	0.74	3.35	4.11
	MZ R^2	8.38 [0.98]	10.16	0.82	4.62	5.34	7.88 [0.49]	10.17	0.77	3.83	4.57	7.54 [0.26]	10.17	0.74	3.30	4.06
$\eta = 1$	SE	8.30 [0.80]	10.16	0.82	4.70	5.42	7.79 [0.35]	10.16	0.77	3.92	4.65	7.45 [0.17]	10.17	0.73	3.40	4.15
	QLIKE	8.45 [0.85]	10.18	0.83	4.55	5.29	7.95 [0.66]	10.19	0.78	3.76	4.52	7.62 [0.39]	10.19	0.75	3.23	4.00
	EL 20	8.57 [0.58]	10.16	0.84	4.43	5.15	8.07 [0.92]	10.17	0.79	3.64	4.38	7.74 [0.59]	10.17	0.76	3.11	3.86
	MZ R^2	8.44 [0.86]	10.17	0.83	4.55	5.28	7.95 [0.65]	10.17	0.78	3.76	4.50	7.62 [0.38]	10.18	0.75	3.22	3.98
$\eta = 4$	SE	8.50 [0.73]	10.17	0.84	4.50	5.23	7.99 [0.73]	10.17	0.79	3.72	4.46	7.65 [0.42]	10.18	0.75	3.20	3.95
	QLIKE	8.55 [0.66]	10.28	0.83	4.46	5.31	8.05 [0.90]	10.30	0.78	3.67	4.54	7.72 [0.61]	10.31	0.75	3.14	4.03
	EL 20	8.70 [0.40]	10.24	0.85	4.31	5.11	8.17 [0.85]	10.25	0.80	3.54	4.37	7.82 [0.81]	10.26	0.76	3.03	3.87
	MZ R^2	8.55 [0.64]	10.19	0.84	4.45	5.21	8.07 [0.94]	10.20	0.79	3.64	4.41	7.76 [0.66]	10.20	0.76	3.09	3.88
$\eta = \infty$	SE	8.83 [0.33]	10.28	0.86	4.18	5.03	8.11 [0.99]	10.29	0.79	3.61	4.48	7.63 [0.53]	10.30	0.74	3.23	4.12
	QLIKE	8.69 [0.51]	10.37	0.84	4.33	5.27	7.97 [0.78]	10.38	0.77	3.75	4.72	7.50 [0.38]	10.39	0.72	3.37	4.35
	EL 20	8.86 [0.30]	10.35	0.86	4.16	5.07	8.16 [0.91]	10.36	0.79	3.57	4.51	7.69 [0.63]	10.37	0.74	3.18	4.14
	MZ R^2	8.37 [0.98]	10.19	0.82	4.63	5.39	7.60 [0.24]	10.20	0.74	4.11	4.89	7.08 [0.05]	10.21	0.69	3.77	4.56

Notes: Average annualized excess mean return (Avg), annualized standard deviation (Std), and Sharpe Ratio (SR). Δ_γ is the annualized fee in percent an investor would be willing to pay for switching to the infeasible oracle portfolio; see Equation (2.4). p-values for two-sided tests of equal returns using Newey-West standard errors with three lags against the benchmark model are reported in brackets. The evaluation period is 2002:M1–2018:M12.

Table 2.6: Low-volatility portfolio returns and factor loadings.

	CAPM		FFC					FF five-factor					
	α	β_{MKT}	α	β_{MKT}	β_{SMB}	β_{HML}	β_{MOM}	α	β_{MKT}	β_{SMB}	β_{HML}	β_{RMW}	β_{CMA}
12m-RV^d	4.38	0.58	4.10	0.63	-0.13	0.08	0.09	2.62	0.67	-0.08	-0.05	0.28	0.23
	[0.00]	[0.00]	[0.00]	[0.00]	[0.01]	[0.15]	[0.01]	[0.02]	[0.00]	[0.14]	[0.46]	[0.00]	[0.04]
SE	4.62	0.61	4.42	0.66	-0.13	0.07	0.07	2.98	0.69	-0.09	-0.06	0.25	0.27
	[0.00]	[0.00]	[0.00]	[0.00]	[0.02]	[0.15]	[0.02]	[0.01]	[0.00]	[0.11]	[0.40]	[0.00]	[0.01]
QLIKE	4.43	0.61	4.18	0.66	-0.11	0.07	0.08	2.84	0.70	-0.07	-0.05	0.24	0.26
	[0.00]	[0.00]	[0.00]	[0.00]	[0.05]	[0.15]	[0.01]	[0.01]	[0.00]	[0.20]	[0.43]	[0.00]	[0.01]
EL 20	4.60	0.61	4.35	0.66	-0.11	0.08	0.08	3.13	0.69	-0.07	-0.04	0.23	0.22
	[0.00]	[0.00]	[0.00]	[0.00]	[0.06]	[0.13]	[0.01]	[0.00]	[0.00]	[0.24]	[0.59]	[0.00]	[0.02]
MZ R^2	4.22	0.60	4.01	0.65	-0.12	0.09	0.07	2.54	0.69	-0.08	-0.05	0.25	0.28
	[0.00]	[0.00]	[0.00]	[0.00]	[0.03]	[0.08]	[0.02]	[0.02]	[0.00]	[0.13]	[0.48]	[0.00]	[0.01]

Notes: Fama-French regressions for the low-volatility portfolio returns in the leading case with $\eta = \infty$ and $\delta = 0.98$. As factors we consider the excess market return *MKT*, the size factor *SMB*, the value factor *HML* in conjunction with the momentum factor *MOM*, or the profitability factor *RMW* and the investment factor *CMA*. Regarding the factor loading coefficients, we report p-values based on two-sided t-tests and Newey-West standard errors in brackets. Excess returns are reported on an annualized scale. The evaluation period is 2002:M1–2018:M12.

2.8 Appendix

2.8.1 Description of time series models

Because months have different numbers of days, all models forecast the 22-day-ahead average realized variance which is then evaluated against the average realized variance in that month. Let $\mathcal{F}_t$ denote the information set up to time t.

HAR-type models

- **HAR**: The HAR model (Corsi, 2009) employs the realized variances directly. In this model, realized variances are regressed on past realized variances aggregated on a daily, weekly, and monthly frequency. The model for forecasting the 22-day-ahead cumulative variance is given by

$$RV_{i,t+1:t+22} = b_0 + b_d RV_{i,t} + b_w RV_{i,t-4:t} + b_m RV_{i,t-21:t} + \eta_{i,t}$$

  with $RV_{i,t+1:t+l} = \sum_{k=1}^{l} RV_{i,t+k}$ and $\mathbf{E}[\eta_{i,t}|\mathcal{F}_{t-1}] = 0$.

- **HAR-SPX:** Now, let $RV_{mkt,t}$ denote the realized variance of the S&P 500 index. Then the HAR-SPX model is the HAR model from above augmented by a HAR model forecast for the market itself,

$$RV_{i,t+1:t+22} = b_0^S + b_d^S RV_{i,t} + b_w^S RV_{i,t-4:t} + b_m^S RV_{i,t-21:t} + b_{mkt}^S \widehat{RV}_{mkt,t+1:t+22|t} + \eta_{i,t}^S$$

  with $\mathbf{E}[\eta_{i,t}^S|\mathcal{F}_{t-1}] = 0$.

- **HAR-LR:** Given that we are only interested in monthly volatility forecast, we employ a long-run version of the HAR model that includes a quarterly and semiannual component:

$$\begin{aligned} RV_{i,t+1:t+22} = b_0^L + & b_d^L RV_{i,t} + b_w^L RV_{i,t-4:t} + b_m^L RV_{i,t-21:t} \\ & + b_q^L RV_{i,t-65:t} + b_s^L RV_{i,t-131:t} + \eta_{i,t}^L \end{aligned}$$

  with $\mathbf{E}[\eta_{i,t}^L|\mathcal{F}_{t-1}] = 0$.

- **HAR-SPX-LR**: As we did in the HAR-SPX, we can als define a HAR-SPX-LR model which employs both the long-run and the market component,

$$\begin{aligned}RV_{i,t+1:t+22} = b_0^{SL} &+ b_d^{SL} RV_{i,t} + b_w^{SL} RV_{i,t-4:t} + b_m^{SL} RV_{i,t-21:t} \\ &+ b_q^{SL} RV_{i,t-65:t} + b_s^{SL} RV_{i,t-131:t} + b_{mkt}^{SL} \widehat{RV}_{mkt,t+1:t+22|t} + \eta_{i,t}^{SL}\end{aligned}$$

with $\mathbf{E}[\eta_{i,t}^{SL}|\mathcal{F}_{t-1}] = 0$.

- **Panel HAR:** The HAR model can also be estimated in a panel if the individual realized variances are demeaned first. Let $\overline{RV}_i$ be the average realized variance of stock i in the estimation period. Then we estimate Panel HAR coefficients

$$RV_{i,t+1:t+22} - \overline{RV}_i = b_d^P(RV_{i,t} - \overline{RV}_i) + b_w^P(RV_{i,t-4:t} - \overline{RV}_i) + b_m^P(RV_{i,t-21:t} - \overline{RV}_i) + \eta_{i,t}^P$$

with $\mathbf{E}[\eta_{i,t}^{P}|\mathcal{F}_{t-1}] = 0$. For forecasting the individual stock's realized variance, we re-add $\overline{RV}_i$ in the end.

- **Panel HAR-LR**: The Panel HAR-LR model is then the long-run analogue of the Panel HAR:

$$\begin{aligned}RV_{i,t+1:t+22} - \overline{RV}_i &= b_d^{PL}(RV_{i,t} - \overline{RV}_i) + b_w^{PL}(RV_{i,t-4:t} - \overline{RV}_i) \\ &+ b_m^{PL}(RV_{i,t-21:t} - \overline{RV}_i) + b_q^{PL}(RV_{i,t-65:t} - \overline{RV}_i) + b_s^{PL}(RV_{i,t-131:t} - \overline{RV}_i) + \eta_{i,t}^{PL}\end{aligned}$$

with $\mathbf{E}[\eta_{i,t}^{PL}|\mathcal{F}_{t-1}] = 0$.

- **HAR-VIX**: All models above are only backward-looking time series models and make no use of expectations on future volatility; for example, those implied by option prices. Hence, we include the squared VIX as a model-free risk-neutral measure of next-month's volatility of market returns,

$$RV_{\,i,t+1:t+22|t} = b_0^V + b_d^V RV_{i,t} + b_w^V RV_{i,t-4:t} + b_m^V RV_{i,t-21:t} + b_{\text{vix}} VIX_t^2 + \eta_{i,t}^V.$$

with $\mathbf{E}[\eta_{i,t}^{V}|\mathcal{F}_{t-1}] = 0$. Bekaert and Hoerova (2014) use the same approach for forecasting aggregate stock market volatility instead of individual stocks. Of course, one could derive individual option-implied volatilities from each stock's option prices but that is beyond the scope of this paper.

- **HAR-VIX-LR**: The HAR-VIX model may also be augmented by our two long-run components:

$$\begin{aligned} RV_{i,t+1:t+22|t} =& b_0^{VL} + b_d^{VL} RV_{i,t} + b_w^{VL} RV_{i,t-4:t} + b_m^{VL} RV_{i,t-21:t} \\ &+ b_q^{VL} RV_{i,t-66:t} + b_s^{VL} RV_{i,t-132:t} + b_{\text{vix}}^{VL} VIX_t^2 + \eta_{i,t}^{VL}. \end{aligned}$$

  with $\mathbf{E}[\eta_{i,t}^{VL}|\mathcal{F}_{t-1}] = 0$.

All HAR models are estimated by ordinary least squares estimation.

GARCH-type models

Let $\varepsilon_{mkt,t}$ and $\varepsilon_{i,t}$ denote the demeaned market and individual stock log returns. Likewise, let $\bar{\sigma}^2_{mkt}$ and $\bar{\sigma}^2_i$ denote the empirical variances of the two in the corresponding estimation sample.

- **GJR-GARCH:** The GARCH specification of Glosten, Jagannathan, and Runkle (1993) of returns $\varepsilon_{i,t} = \sqrt{h_{i,t}^{GJR}} Z_{i,t}^{GJR}$, $Z_{i,t}^{GJR} \sim \mathcal{D}(0,1)$, is given by

$$h_{i,t}^{GJR} = (1 - \alpha_i^{GJR} - \beta_i^{GJR} - \gamma_i^{GJR}/2)\bar{\sigma}_i^2 + \alpha_i^{GJR}\varepsilon_{i,t-1}^2 + \gamma_i^{GJR}\mathbb{1}_{\{\varepsilon_{i,t-1}<0\}}\varepsilon_{i,t-1}^2 + \beta_i^{GJR} h_{i,t-1}^{GJR}.$$

  We determine the rolling-window coefficients by quasi-maximum-likelihood estimation (QMLE).

- **Panel GJR-GARCH:** Instead of estimating the GARCH coefficients for every stock separately, we can estimate a Panel GJR-GARCH in which

$$\begin{aligned} h_{i,t}^{PGJR} =& (1 - \alpha^{PGJR} - \beta^{PGJR} - \gamma^{PGJR}/2)\bar{\sigma}_i^2 + \alpha^{PGJR}\varepsilon_{i,t-1}^2 \\ &+ \gamma^{PGJR}\mathbb{1}_{\{\varepsilon_{i,t-1}<0\}}\varepsilon_{i,t-1}^2 + \beta^{PGJR} h_{i,t-1}^{PGJR}. \end{aligned}$$

  Under the assumption of the innovation terms being independent, the Panel GJR-GARCH is estimated via QMLE by summing up the individual log-likelihoods.

- **Factor GARCH:** In this model introduced by Engle, Ng, and Rothschild (1990), the market return is modeled as a GJR-GARCH,

$$\varepsilon_{mkt,t} = \sqrt{h^{CG}_{mkt,t}} Z^{CG}_{mkt,t},$$

with $Z_{mkt,t} \sim \mathcal{D}(0,1)$ and

$$\begin{aligned} h_{mkt,t} = (1 - \alpha^{CG}_{mkt} - \beta^{CG}_{mkt} - \gamma^{CG}_{mkt}/2)\bar{\sigma}^2_{mkt} + \alpha^{CG}_{mkt}\varepsilon^2_{mkt,t-1} \\ + \gamma^{CG}_{mkt}\mathbb{1}_{\{\varepsilon_{mkt,t-1}<0\}} r^2_{mkt,t-1} + \beta^{CG}_{mkt} h^{CG}_{mkt,t-1}. \end{aligned}$$

The individual demeaned stock return is given by

$$\varepsilon_{i,t} = \beta^{CG}_i r_{mkt,t} + \eta^{CG}_{i,t} = \beta^{CG}_i r_{mkt,t} + \sqrt{h^{CG}_{i,t}} Z^{CG}_{i,t}$$

with $Z^{CG}_{i,t} \sim \mathcal{D}(0,1)$ and

$$h^{CG}_{i,t} = (1 - \alpha^{CG}_i - \beta^{CG}_i)\bar{\omega}_i + \alpha^{CG}_i \eta^2_{i,t-1} + \beta_i h^{CG}_{i,t-1},$$

where $\bar{\omega}_i$ denotes the empirical variance of the stock-specific CAPM residuals. Under the assumption of independence of $Z^{CG}_{mkt,t}$ and $Z^{CG}_{i,t}$, the forecast of the individual stock's conditional variance is given by

$$\left(\beta^{CG}_i\right)^2 h^{CG}_{mkt,t+1:t+22|t} + h^{CG}_{i,t+1:t+22|t}$$

where $h^{CG}_{mkt,t+1:t+22|t}$ and $h^{CG}_{i,t+1:t+22|t}$ are the cumulated daily GARCH forecasts. The β^{CG}_is are estimated separately for each stock in the respective rolling window as well as the GARCH models for the market and the CAPM-residuals.

- The **Factor GARCH-MIDAS** model is the same as the CAPM GARCH model but the market return is now given by a GARCH-MIDAS model. It includes either the VIX, changes in housing starts, or the term spread as a covariate and estimation has been carried out using QMLE using the R-package *mfGARCH* by Kleen (2018).

  More specifically, the standardized demeaned market return $\varepsilon_{mkt,t}$ is now modeled as

$$\frac{\varepsilon_{mkt,t}}{\sqrt{\tau_t}} = \sqrt{g_{mkt,t}} Z_{mkt,t},$$

where τ_t is specified as a function of a monthly explanatory variable X_m, $g_{mkt,t}$ follows a daily GARCH equation, and $Z_{mkt,t}$ is an *i.i.d.* innovation process with mean zero and variance one. The short-term component is assumed to follow a mean-reverting unit-variance GJR-GARCH process:

$$\begin{aligned} g_{mkt,t} = {} & (1 - \alpha^{CGM} - \gamma^{CGM}/2 - \beta^{CGM}) \\ & + \left(\alpha^{CGM} + \gamma^{CGM} \mathbb{1}_{\{\varepsilon_{mkt,t-1} < 0\}}\right) \frac{\varepsilon^2_{mkt,t-1}}{\tau_m} + \beta^{CGM} g_{mkt,t-1}. \end{aligned}$$

The long-term component τ_m in month m is given by

$$\tau_m = \exp\left(m^{CGM} + \theta^{CGM} \sum_{l=1}^{K} \varphi_l(w_1^{CGM}, w_2^{CGM}) X_{m-l} \right).$$

where the weights $\varphi_l(w_1, w_2) \geq 0$ are parameterized via the Beta weighting scheme

$$\varphi_l(w_1, w_2) = \frac{(l/(K+1))^{w_1-1} \cdot (1 - l/(K+1))^{w_2-1}}{\sum_{j=1}^{K} (j/(K+1))^{w_1-1} \cdot (1 - j/(K+1))^{w_2-1}}. \tag{2.5}$$

In our versions with either changes in housing starts or the term spread as the explanatory variable X_m, we choose $K = 36$. In case of the VIX, we choose $K = 3$. For more details see Conrad and Kleen (2020). We name our Factor GARCH-MIDAS models accordingly to the covariate employed: *Factor GARCH-VIX*, *Factor GARCH-ΔHous*, and *Factor GARCH-TS*.

- **Realized GARCH:** As a generalization of the GARCH model, we employ the Realized GARCH model (Hansen et al., 2012). Here, the conditional variance of the returns $\varepsilon_t = \sqrt{\sigma_t^{RG}} Z_t^{RG}$, $Z_t^{RG} \overset{i.i.d.}{\sim} \mathcal{D}(0, 1)$ at day t is modeled as

$$\log \sigma_t^{RG} = \omega^{RG} + \alpha^{RG} \log RV_{t-1}^{int} + \beta^{RG} \log \sigma_{t-1}^{RG}$$

and the realized measure RV_t^{int} based on intraday returns only as

$$\log RV_t^{int} = \xi^{RG} + \delta^{RG} \log \sigma_t^{RG} + \eta_1^{RG} Z_t^{RG} + \eta_2^{RG} \left(\left(Z_t^{RG} \right)^2 - 1 \right) + u_t^{RG}$$

with $u_t^{RG} \overset{i.i.d.}{\sim} \mathcal{N}(0, \lambda^{RG})$. The innovations Z_t^{RG} and u_t^{RG} are independent. The estimation of the Realized GARCH model and the forecast computation by simulation is carried out using the R-package *rugarch* (Ghalanos, 2018).

- **Multiplicative Error Model:** The Multiplicative Error Model (MEM) by Engle and Gallo (2006) employs as the dependent variable not (demeaned) returns but the realized measure itself, $\sqrt{RV_{i,t}} = h_{i,t}^{MEM} Z_{i,t}^{MEM}$, $Z_{i,t}^{MEM} \sim \mathcal{D}(0,1)$, and

$$h_{i,t}^{MEM} = (1 - \alpha_i^{MEM} - \beta_i^{MEM})\overline{RV}_i + \alpha_i^{MEM} RV_{i,t-1} + \beta_i^{MEM} h_{i,t-1}^{MEM}$$

with $\overline{RV}_i$ being the average $RV_{i,t}$ in the corresponding rolling estimation sample.

- **Panel MEM:** As in the Panel GARCH, we can estimate one parameter vector for all stocks in a Panel MEM model by summing up the log-likelihoods with respect to all centered conditional variance equations jointly,

$$h_{i,t}^{PMEM} = (1 - \alpha^{PMEM} - \beta^{PMEM})\overline{RV}_i + \alpha^{PMEM} RV_{i,t-1} + \beta^{PMEM} h_{i,t-1}^{PMEM}.$$

MIDAS-type models

- **MIDAS:** The class of MIDAS models was introduced by Ghysels, Santa-Clara, and Valkanov (2004, 2005, 2006) which are very flexible distributed lag models that potentially employ data sampled on different frequencies (see the CAPM GARCH-MIDAS above). In our case, the model is defined as

$$RV_{i,t+1:t+22|t} - \overline{RV}_i = \theta_i^M \sum_{l=0}^{K-1} \varphi_l(1, w_{i,2}^M) \cdot (RV_{i,t-l} - \overline{RV}_i) + \eta_{i,t}^M.$$

The weighting scheme is a Beta weighting scheme as in Equation (2.5) with $w_1 = 1$ and we choose $K = 132$ to match the long-run HAR models. We assume $\mathbf{E}[\eta_{i,t}^M | \mathcal{F}_{t-1}] = 0$. The parameters are obtained by minimizing the squared residuals.

- **Panel MIDAS:** Similar to our other panel variations for HAR and GARCH models, we include a Panel MIDAS by restricting the scaling parameter θ_i^M and the weighting parameter $w_{i,2}^M$ to be the same for all stocks,

$$RV_{i,t+1:t+22|t} - \overline{RV}_i = \theta^{PM} \sum_{l=0}^{K-1} \varphi_l(1, w_2^{PM}) \cdot (RV_{i,t-l} - \overline{RV}_i) + \eta_{i,t}^{PM}.$$

We assume $\mathbf{E}[\eta_{i,t}^{PM}|\mathcal{F}_{t-1}] = 0$. This is again estimated by minimizing the squared residuals.

Riskmetrics

Our Riskmetrics forecasts are based either on monthly (indexed by m) or daily data (indexed by t). In total we employ four different versions. The first is *RM monthly, 12 months* and the forecasts are given by

$$RV^d_{m+1|m} = \frac{1}{\sum_{k=0}^{K-1} \lambda^k} \sum_{k=0}^{K-1} \lambda^k RV^d_{m-k}$$

with $K = 12$ and RV^d_m being the realized variance in month m based on squared daily returns. *RM monthly, 6 months* is the same but with $K = 6$. *RM daily, 12 months, and RM daily, 6 months* are similar but they use daily squared returns on the right hand side with the corresponding number of lags to match the data of the monthly RM models. We choose $\lambda = 0.97$ because we target the monthly horizon.

All models are reestimated at the end of each month. In a handful of cases, the forecast is unreasonable (e.g., negative for some stocks in the Panel HAR model). Thus, we apply a rolling "sanity filter" which truncates forecasts by the 0.1%- and 99.9%-quantile of cross-sectional monthly RVs in the estimation window.

2.8.2 Returns for additional combinations of η and δ

In this section, we discuss the possible alternative choices for δ in our empirical analysis. As such, we report the average returns for our loss-based strategies in Table 2.7–2.10. Moreover, we report alternative values for the proportional transaction costs c. In the case of SE and QLIKE, we observe the highest average return for $\eta = \infty$ and $\delta = 1$; that is, only considering the best model and putting equal weight on all past cross-sectional forecast errors. For the EL 20, the ex-post optimal combination without transaction costs is $\eta = \infty$ and $\delta = 0.9$. Among the three loss function, EL 20 has the largest increase in average returns by up to more than 1 percentage point in the case of $\eta = \infty$. Interestingly, for the MZ R^2 the highest returns are observed for $\eta = \infty$ and $\delta = 0$.

Table 2.7: Average returns of SE-based portfolios.

		δ								
c	η	0	0.6	0.8	0.9	0.94	0.98	0.99	0.999	1
0	0	8.43	—	—	—	—	—	—	—	—
		(0.14)								
	0.5	8.37	8.32	8.41	8.40	8.37	8.33	8.32	8.32	8.32
		(-0.04)	(-0.20)	(0.07)	(0.04)	(-0.05)	(-0.18)	(-0.21)	(-0.19)	(-0.19)
	1	8.29	8.27	8.31	8.37	8.33	8.30	8.30	8.34	8.34
		(-0.29)	(-0.34)	(-0.23)	(-0.05)	(-0.17)	(-0.25)	(-0.25)	(-0.14)	(-0.15)
	2	8.41	8.27	8.20	8.24	8.35	8.34	8.35	8.34	8.34
		(0.08)	(-0.34)	(-0.54)	(-0.42)	(-0.10)	(-0.12)	(-0.11)	(-0.12)	(-0.13)
	4	8.34	8.31	8.30	8.44	8.45	8.50	8.47	8.46	8.46
		(-0.14)	(-0.21)	(-0.26)	(0.17)	(0.22)	(0.34)	(0.26)	(0.23)	(0.23)
	10	8.42	8.39	8.32	8.49	8.53	8.53	8.50	8.50	8.49
		(0.11)	(0.03)	(-0.20)	(0.35)	(0.44)	(0.44)	(0.35)	(0.32)	(0.32)
	Inf	8.79	8.53	8.48	8.76	8.93	8.83	8.88	9.12	9.12
		(1.01)	(0.34)	(0.23)	(0.90)	(1.26)	(0.97)	(1.09)	(1.65)	(1.65)
5	0	8.26	—	—	—	—	—	—	—	—
		(-0.10)								
	0.5	8.20	8.15	8.24	8.23	8.20	8.16	8.15	8.15	8.15
		(-0.27)	(-0.43)	(-0.15)	(-0.19)	(-0.28)	(-0.41)	(-0.45)	(-0.43)	(-0.43)
	1	8.12	8.10	8.14	8.20	8.16	8.13	8.13	8.17	8.17
		(-0.52)	(-0.57)	(-0.46)	(-0.27)	(-0.40)	(-0.48)	(-0.48)	(-0.37)	(-0.37)
	2	8.24	8.10	8.04	8.08	8.18	8.17	8.18	8.17	8.17
		(-0.17)	(-0.57)	(-0.75)	(-0.63)	(-0.32)	(-0.34)	(-0.32)	(-0.34)	(-0.34)
	4	8.15	8.14	8.13	8.27	8.28	8.33	8.30	8.29	8.29
		(-0.42)	(-0.44)	(-0.48)	(-0.07)	(-0.02)	(0.11)	(0.02)	(-0.01)	(-0.01)
	10	8.20	8.22	8.15	8.32	8.36	8.36	8.33	8.32	8.32
		(-0.26)	(-0.22)	(-0.45)	(0.11)	(0.20)	(0.20)	(0.12)	(0.10)	(0.09)
	Inf	8.46	8.26	8.22	8.51	8.69	8.59	8.65	8.89	8.89
		(0.44)	(-0.08)	(-0.18)	(0.54)	(0.91)	(0.65)	(0.77)	(1.33)	(1.33)
10	0	8.09	—	—	—	—	—	—	—	—
		(-0.33)								
	0.5	8.03	7.98	8.07	8.06	8.03	7.99	7.98	7.98	7.98
		(-0.50)	(-0.66)	(-0.38)	(-0.42)	(-0.51)	(-0.65)	(-0.68)	(-0.66)	(-0.66)
	1	7.95	7.93	7.97	8.03	7.99	7.96	7.96	8.00	8.00
		(-0.76)	(-0.79)	(-0.68)	(-0.49)	(-0.63)	(-0.71)	(-0.70)	(-0.59)	(-0.60)
	2	8.06	7.93	7.87	7.91	8.01	8.00	8.01	8.01	8.00
		(-0.41)	(-0.80)	(-0.97)	(-0.85)	(-0.53)	(-0.55)	(-0.54)	(-0.55)	(-0.55)
	4	7.97	7.97	7.96	8.10	8.11	8.16	8.13	8.12	8.12
		(-0.69)	(-0.66)	(-0.70)	(-0.31)	(-0.26)	(-0.12)	(-0.21)	(-0.24)	(-0.24)
	10	7.99	8.05	7.97	8.16	8.19	8.18	8.16	8.15	8.15
		(-0.62)	(-0.48)	(-0.69)	(-0.13)	(-0.03)	(-0.04)	(-0.11)	(-0.13)	(-0.14)
	Inf	8.14	8.00	7.96	8.27	8.44	8.35	8.41	8.66	8.66
		(-0.14)	(-0.50)	(-0.59)	(0.18)	(0.57)	(0.33)	(0.45)	(1.02)	(1.02)
15	0	7.92	—	—	—	—	—	—	—	—
		(-0.57)								
	0.5	7.86	7.81	7.90	7.89	7.86	7.82	7.81	7.81	7.81
		(-0.73)	(-0.88)	(-0.61)	(-0.66)	(-0.74)	(-0.88)	(-0.92)	(-0.89)	(-0.89)
	1	7.77	7.76	7.80	7.86	7.82	7.79	7.79	7.83	7.83
		(-0.99)	(-1.02)	(-0.90)	(-0.71)	(-0.85)	(-0.93)	(-0.93)	(-0.82)	(-0.82)
	2	7.88	7.76	7.70	7.74	7.84	7.83	7.84	7.84	7.83
		(-0.66)	(-1.03)	(-1.18)	(-1.07)	(-0.75)	(-0.77)	(-0.75)	(-0.77)	(-0.77)
	4	7.78	7.81	7.80	7.93	7.94	7.99	7.96	7.95	7.95
		(-0.96)	(-0.88)	(-0.93)	(-0.55)	(-0.50)	(-0.35)	(-0.44)	(-0.48)	(-0.48)
	10	7.77	7.87	7.80	7.99	8.02	8.01	7.99	7.98	7.98
		(-0.98)	(-0.73)	(-0.94)	(-0.37)	(-0.27)	(-0.27)	(-0.34)	(-0.35)	(-0.36)
	Inf	7.82	7.73	7.70	8.03	8.20	8.11	8.17	8.43	8.43
		(-0.73)	(-0.92)	(-1.00)	(-0.19)	(0.22)	(0.01)	(0.14)	(0.72)	(0.72)
20	0	7.75	—	—	—	—	—	—	—	—
		(-0.80)								
	0.5	7.69	7.63	7.73	7.72	7.69	7.65	7.64	7.64	7.64
		(-0.96)	(-1.11)	(-0.83)	(-0.89)	(-0.97)	(-1.11)	(-1.15)	(-1.12)	(-1.12)
	1	7.60	7.59	7.63	7.69	7.65	7.62	7.62	7.66	7.66
		(-1.23)	(-1.24)	(-1.12)	(-0.93)	(-1.08)	(-1.16)	(-1.15)	(-1.04)	(-1.05)
	2	7.71	7.59	7.53	7.57	7.67	7.66	7.67	7.67	7.66
		(-0.90)	(-1.25)	(-1.40)	(-1.29)	(-0.96)	(-0.98)	(-0.97)	(-0.98)	(-0.98)
	4	7.60	7.64	7.63	7.76	7.77	7.82	7.79	7.78	7.78
		(-1.23)	(-1.10)	(-1.15)	(-0.79)	(-0.74)	(-0.58)	(-0.67)	(-0.71)	(-0.71)
	10	7.55	7.70	7.62	7.82	7.85	7.84	7.82	7.81	7.81
		(-1.33)	(-0.98)	(-1.17)	(-0.61)	(-0.51)	(-0.50)	(-0.56)	(-0.58)	(-0.59)
	Inf	7.49	7.47	7.44	7.78	7.96	7.87	7.93	8.19	8.19
		(-1.33)	(-1.33)	(-1.40)	(-0.55)	(-0.12)	(-0.31)	(-0.18)	(0.41)	(0.41)
25	0	7.58	—	—	—	—	—	—	—	—
		(-1.03)								
	0.5	7.52	7.46	7.56	7.55	7.52	7.48	7.47	7.47	7.47
		(-1.19)	(-1.33)	(-1.06)	(-1.12)	(-1.20)	(-1.34)	(-1.38)	(-1.35)	(-1.35)
	1	7.43	7.42	7.46	7.52	7.48	7.45	7.45	7.49	7.49
		(-1.46)	(-1.47)	(-1.34)	(-1.15)	(-1.30)	(-1.38)	(-1.37)	(-1.26)	(-1.27)
	2	7.53	7.42	7.36	7.40	7.50	7.49	7.50	7.50	7.49
		(-1.15)	(-1.48)	(-1.61)	(-1.51)	(-1.18)	(-1.20)	(-1.18)	(-1.19)	(-1.19)
	4	7.41	7.47	7.46	7.59	7.60	7.65	7.62	7.61	7.61
		(-1.49)	(-1.32)	(-1.36)	(-1.02)	(-0.97)	(-0.80)	(-0.90)	(-0.94)	(-0.94)
	10	7.34	7.52	7.45	7.65	7.67	7.67	7.65	7.64	7.64
		(-1.69)	(-1.22)	(-1.41)	(-0.84)	(-0.74)	(-0.73)	(-0.79)	(-0.80)	(-0.81)
	Inf	7.17	7.20	7.18	7.54	7.71	7.63	7.69	7.96	7.96
		(-1.93)	(-1.75)	(-1.80)	(-0.91)	(-0.47)	(-0.62)	(-0.49)	(0.10)	(0.10)

Notes: Average annualized excess SE-based mean returns for different combinations of η, δ, and proportional transaction costs c. Weights are given by Equation (2.3). t-statistics for two-sided tests of equal returns using Newey-West standard errors with three lags against the benchmark model 12m-RV^d are reported in parentheses. The evaluation period is 2002:M1–2018:M12.

Table 2.8: Average returns of QLIKE-based portfolios.

		δ								
c	η	0	0.6	0.8	0.9	0.94	0.98	0.99	0.999	1
0	0	8.43 (0.14)	—	—	—	—	—	—	—	—
	0.5	8.37 (-0.06)	8.41 (0.07)	8.42 (0.10)	8.42 (0.10)	8.40 (0.03)	8.40 (0.05)	8.38 (-0.02)	8.40 (0.04)	8.40 (0.05)
	1	8.40 (0.04)	8.41 (0.09)	8.44 (0.17)	8.45 (0.20)	8.46 (0.23)	8.45 (0.20)	8.39 (0.02)	8.44 (0.15)	8.44 (0.15)
	2	8.48 (0.29)	8.36 (-0.07)	8.34 (-0.13)	8.41 (0.08)	8.36 (-0.06)	8.42 (0.10)	8.45 (0.19)	8.44 (0.15)	8.45 (0.19)
	4	8.42 (0.09)	8.40 (0.03)	8.47 (0.24)	8.47 (0.24)	8.59 (0.54)	8.55 (0.44)	8.57 (0.49)	8.57 (0.48)	8.59 (0.54)
	10	8.44 (0.15)	8.60 (0.56)	8.46 (0.21)	8.48 (0.27)	8.58 (0.54)	8.89 (1.37)	8.82 (1.22)	8.81 (1.18)	8.80 (1.14)
	Inf	8.33 (-0.13)	8.65 (0.70)	8.61 (0.61)	8.57 (0.51)	8.54 (0.40)	8.69 (0.66)	8.90 (1.08)	8.94 (1.18)	8.94 (1.18)
5	0	8.26 (-0.10)	—	—	—	—	—	—	—	—
	0.5	8.20 (-0.28)	8.24 (-0.15)	8.25 (-0.11)	8.25 (-0.11)	8.23 (-0.19)	8.23 (-0.17)	8.21 (-0.24)	8.23 (-0.18)	8.24 (-0.17)
	1	8.23 (-0.18)	8.25 (-0.13)	8.28 (-0.04)	8.29 (-0.01)	8.30 (0.01)	8.28 (-0.02)	8.22 (-0.20)	8.27 (-0.06)	8.27 (-0.06)
	2	8.31 (0.05)	8.19 (-0.28)	8.17 (-0.32)	8.25 (-0.11)	8.20 (-0.26)	8.26 (-0.09)	8.29 (-0.00)	8.28 (-0.04)	8.29 (-0.01)
	4	8.22 (-0.19)	8.22 (-0.19)	8.30 (0.03)	8.30 (0.04)	8.42 (0.34)	8.39 (0.25)	8.41 (0.30)	8.40 (0.29)	8.42 (0.35)
	10	8.22 (-0.21)	8.41 (0.31)	8.27 (-0.05)	8.30 (0.01)	8.40 (0.29)	8.71 (1.13)	8.64 (0.97)	8.63 (0.94)	8.62 (0.90)
	Inf	8.03 (-0.63)	8.39 (0.25)	8.35 (0.17)	8.31 (0.04)	8.28 (-0.02)	8.45 (0.34)	8.66 (0.77)	8.70 (0.87)	8.70 (0.87)
10	0	8.09 (-0.33)	—	—	—	—	—	—	—	—
	0.5	8.03 (-0.50)	8.07 (-0.37)	8.08 (-0.33)	8.08 (-0.33)	8.06 (-0.41)	8.06 (-0.39)	8.04 (-0.46)	8.06 (-0.40)	8.07 (-0.39)
	1	8.06 (-0.40)	8.08 (-0.34)	8.11 (-0.25)	8.12 (-0.22)	8.13 (-0.20)	8.12 (-0.23)	8.06 (-0.41)	8.10 (-0.27)	8.10 (-0.27)
	2	8.13 (-0.19)	8.02 (-0.49)	8.00 (-0.52)	8.08 (-0.30)	8.03 (-0.45)	8.09 (-0.27)	8.13 (-0.19)	8.11 (-0.23)	8.13 (-0.20)
	4	8.03 (-0.47)	8.05 (-0.40)	8.13 (-0.18)	8.14 (-0.17)	8.25 (0.14)	8.22 (0.06)	8.24 (0.11)	8.24 (0.10)	8.26 (0.16)
	10	7.99 (-0.56)	8.22 (0.05)	8.08 (-0.31)	8.11 (-0.24)	8.22 (0.05)	8.53 (0.90)	8.46 (0.73)	8.45 (0.71)	8.44 (0.66)
	Inf	7.73 (-1.13)	8.12 (-0.19)	8.10 (-0.26)	8.04 (-0.43)	8.03 (-0.44)	8.21 (0.03)	8.42 (0.47)	8.47 (0.57)	8.47 (0.57)
15	0	7.92 (-0.57)	—	—	—	—	—	—	—	—
	0.5	7.86 (-0.72)	7.90 (-0.59)	7.91 (-0.54)	7.92 (-0.54)	7.89 (-0.63)	7.90 (-0.61)	7.87 (-0.68)	7.90 (-0.62)	7.90 (-0.61)
	1	7.89 (-0.63)	7.92 (-0.55)	7.94 (-0.46)	7.95 (-0.43)	7.96 (-0.41)	7.95 (-0.45)	7.89 (-0.62)	7.94 (-0.48)	7.94 (-0.48)
	2	7.96 (-0.44)	7.85 (-0.70)	7.84 (-0.71)	7.92 (-0.48)	7.87 (-0.64)	7.93 (-0.46)	7.96 (-0.38)	7.95 (-0.42)	7.96 (-0.39)
	4	7.84 (-0.75)	7.87 (-0.61)	7.96 (-0.39)	7.97 (-0.37)	8.08 (-0.05)	8.05 (-0.13)	8.07 (-0.08)	8.07 (-0.08)	8.09 (-0.03)
	10	7.76 (-0.91)	8.02 (-0.20)	7.89 (-0.56)	7.92 (-0.50)	8.03 (-0.19)	8.35 (0.66)	8.28 (0.49)	8.27 (0.47)	8.26 (0.42)
	Inf	7.43 (-1.63)	7.86 (-0.64)	7.85 (-0.70)	7.78 (-0.89)	7.77 (-0.86)	7.97 (-0.27)	8.18 (0.16)	8.23 (0.26)	8.23 (0.26)
20	0	7.75 (-0.80)	—	—	—	—	—	—	—	—
	0.5	7.70 (-0.94)	7.74 (-0.80)	7.75 (-0.75)	7.75 (-0.75)	7.72 (-0.84)	7.73 (-0.82)	7.71 (-0.89)	7.73 (-0.84)	7.73 (-0.83)
	1	7.72 (-0.85)	7.75 (-0.76)	7.78 (-0.67)	7.79 (-0.64)	7.80 (-0.62)	7.79 (-0.66)	7.73 (-0.83)	7.77 (-0.69)	7.77 (-0.69)
	2	7.79 (-0.68)	7.68 (-0.90)	7.67 (-0.90)	7.75 (-0.67)	7.70 (-0.84)	7.77 (-0.64)	7.80 (-0.56)	7.79 (-0.61)	7.80 (-0.58)
	4	7.65 (-1.02)	7.70 (-0.81)	7.78 (-0.60)	7.80 (-0.58)	7.92 (-0.25)	7.89 (-0.32)	7.91 (-0.26)	7.91 (-0.27)	7.93 (-0.21)
	10	7.54 (-1.27)	7.83 (-0.46)	7.70 (-0.82)	7.74 (-0.75)	7.85 (-0.43)	8.17 (0.43)	8.10 (0.25)	8.09 (0.23)	8.08 (0.18)
	Inf	7.13 (-2.13)	7.59 (-1.08)	7.60 (-1.13)	7.52 (-1.36)	7.51 (-1.28)	7.74 (-0.58)	7.94 (-0.14)	7.99 (-0.04)	7.99 (-0.04)
25	0	7.58 (-1.03)	—	—	—	—	—	—	—	—
	0.5	7.53 (-1.15)	7.57 (-1.02)	7.58 (-0.96)	7.58 (-0.97)	7.56 (-1.06)	7.56 (-1.04)	7.54 (-1.11)	7.56 (-1.05)	7.56 (-1.04)
	1	7.55 (-1.06)	7.58 (-0.97)	7.61 (-0.87)	7.62 (-0.85)	7.63 (-0.83)	7.62 (-0.87)	7.56 (-1.04)	7.61 (-0.90)	7.61 (-0.90)
	2	7.61 (-0.91)	7.52 (-1.11)	7.50 (-1.09)	7.59 (-0.85)	7.54 (-1.03)	7.60 (-0.82)	7.64 (-0.75)	7.62 (-0.79)	7.63 (-0.76)
	4	7.46 (-1.30)	7.52 (-1.02)	7.61 (-0.81)	7.63 (-0.78)	7.75 (-0.44)	7.72 (-0.50)	7.74 (-0.45)	7.74 (-0.45)	7.76 (-0.40)
	10	7.31 (-1.62)	7.64 (-0.71)	7.51 (-1.07)	7.55 (-1.00)	7.67 (-0.67)	7.99 (0.20)	7.92 (0.02)	7.91 (0.00)	7.90 (-0.05)
	Inf	6.83 (-2.63)	7.32 (-1.52)	7.34 (-1.56)	7.26 (-1.82)	7.25 (-1.70)	7.50 (-0.89)	7.70 (-0.45)	7.75 (-0.35)	7.75 (-0.35)

Notes: See Table 2.7 but for QLIKE-based portfolios.

Table 2.9: Average returns of EL 20-based portfolios

		δ								
c	η	0	0.6	0.8	0.9	0.94	0.98	0.99	0.999	1
0	0	8.43 (0.14)	—	—	—	—	—	—	—	—
	0.5	8.34 (-0.14)	8.36 (-0.09)	8.31 (-0.25)	8.37 (-0.05)	8.39 (0.00)	8.33 (-0.16)	8.32 (-0.20)	8.30 (-0.26)	8.30 (-0.26)
	1	8.55 (0.51)	8.38 (-0.02)	8.47 (0.25)	8.48 (0.28)	8.53 (0.44)	8.57 (0.55)	8.55 (0.50)	8.57 (0.56)	8.57 (0.57)
	2	8.28 (-0.31)	8.36 (-0.08)	8.42 (0.08)	8.50 (0.32)	8.54 (0.44)	8.61 (0.60)	8.61 (0.61)	8.61 (0.63)	8.61 (0.62)
	4	8.33 (-0.17)	8.43 (0.11)	8.42 (0.08)	8.60 (0.59)	8.67 (0.79)	8.70 (0.85)	8.69 (0.82)	8.75 (0.99)	8.73 (0.94)
	10	7.99 (-1.13)	8.45 (0.16)	8.63 (0.66)	8.73 (0.94)	8.78 (1.08)	8.77 (1.02)	8.77 (1.02)	8.80 (1.08)	8.78 (1.04)
	Inf	7.72 (-1.72)	8.70 (0.79)	8.79 (0.97)	8.90 (1.22)	8.88 (1.14)	8.86 (1.03)	8.89 (1.08)	8.83 (0.97)	8.83 (0.97)
5	0	8.26 (-0.10)	—	—	—	—	—	—	—	—
	0.5	8.18 (-0.38)	8.19 (-0.31)	8.14 (-0.47)	8.20 (-0.27)	8.22 (-0.21)	8.16 (-0.38)	8.15 (-0.41)	8.13 (-0.47)	8.13 (-0.48)
	1	8.38 (0.28)	8.22 (-0.23)	8.30 (0.04)	8.32 (0.07)	8.37 (0.23)	8.40 (0.33)	8.38 (0.28)	8.40 (0.34)	8.41 (0.36)
	2	8.11 (-0.54)	8.19 (-0.27)	8.25 (-0.11)	8.34 (0.13)	8.38 (0.24)	8.44 (0.40)	8.44 (0.41)	8.45 (0.43)	8.45 (0.42)
	4	8.14 (-0.46)	8.26 (-0.09)	8.25 (-0.12)	8.43 (0.38)	8.50 (0.58)	8.52 (0.63)	8.51 (0.60)	8.58 (0.76)	8.56 (0.71)
	10	7.76 (-1.54)	8.26 (-0.08)	8.45 (0.42)	8.54 (0.69)	8.59 (0.82)	8.57 (0.73)	8.57 (0.73)	8.59 (0.79)	8.58 (0.76)
	Inf	7.40 (-2.33)	8.45 (0.40)	8.57 (0.65)	8.67 (0.89)	8.65 (0.83)	8.62 (0.72)	8.65 (0.78)	8.60 (0.66)	8.60 (0.66)
10	0	8.09 (-0.33)	—	—	—	—	—	—	—	—
	0.5	8.01 (-0.62)	8.03 (-0.54)	7.97 (-0.70)	8.04 (-0.49)	8.05 (-0.43)	8.00 (-0.60)	7.98 (-0.63)	7.97 (-0.69)	7.96 (-0.69)
	1	8.21 (0.06)	8.05 (-0.44)	8.14 (-0.17)	8.15 (-0.14)	8.20 (0.01)	8.23 (0.11)	8.22 (0.07)	8.24 (0.13)	8.24 (0.14)
	2	7.93 (-0.77)	8.03 (-0.46)	8.09 (-0.30)	8.18 (-0.06)	8.22 (0.05)	8.28 (0.21)	8.28 (0.22)	8.28 (0.23)	8.28 (0.22)
	4	7.94 (-0.75)	8.08 (-0.29)	8.08 (-0.32)	8.26 (0.17)	8.33 (0.37)	8.35 (0.41)	8.34 (0.38)	8.40 (0.54)	8.38 (0.49)
	10	7.52 (-1.95)	8.08 (-0.32)	8.26 (0.18)	8.36 (0.44)	8.40 (0.55)	8.37 (0.45)	8.37 (0.45)	8.39 (0.51)	8.38 (0.47)
	Inf	7.07 (-2.94)	8.20 (0.00)	8.34 (0.33)	8.44 (0.57)	8.42 (0.53)	8.39 (0.42)	8.42 (0.48)	8.36 (0.35)	8.36 (0.35)
15	0	7.92 (-0.57)	—	—	—	—	—	—	—	—
	0.5	7.84 (-0.85)	7.86 (-0.77)	7.81 (-0.92)	7.87 (-0.70)	7.89 (-0.64)	7.83 (-0.81)	7.82 (-0.84)	7.80 (-0.90)	7.80 (-0.91)
	1	8.05 (-0.17)	7.89 (-0.65)	7.97 (-0.38)	7.99 (-0.35)	8.04 (-0.20)	8.07 (-0.10)	8.06 (-0.14)	8.07 (-0.09)	8.08 (-0.07)
	2	7.76 (-1.00)	7.86 (-0.65)	7.92 (-0.48)	8.01 (-0.24)	8.05 (-0.14)	8.11 (0.02)	8.11 (0.02)	8.11 (0.03)	8.11 (0.03)
	4	7.75 (-1.04)	7.91 (-0.48)	7.91 (-0.52)	8.09 (-0.04)	8.16 (0.16)	8.17 (0.19)	8.16 (0.16)	8.22 (0.32)	8.20 (0.27)
	10	7.29 (-2.35)	7.89 (-0.55)	8.08 (-0.06)	8.17 (0.18)	8.21 (0.29)	8.17 (0.17)	8.16 (0.16)	8.19 (0.22)	8.17 (0.19)
	Inf	6.75 (-3.54)	7.95 (-0.39)	8.11 (0.01)	8.21 (0.25)	8.20 (0.22)	8.16 (0.12)	8.18 (0.17)	8.12 (0.04)	8.12 (0.04)
20	0	7.75 (-0.80)	—	—	—	—	—	—	—	—
	0.5	7.68 (-1.08)	7.70 (-0.99)	7.64 (-1.14)	7.70 (-0.92)	7.72 (-0.86)	7.66 (-1.02)	7.65 (-1.06)	7.63 (-1.12)	7.63 (-1.12)
	1	7.88 (-0.40)	7.73 (-0.86)	7.81 (-0.59)	7.82 (-0.56)	7.87 (-0.41)	7.90 (-0.32)	7.89 (-0.36)	7.91 (-0.30)	7.92 (-0.28)
	2	7.58 (-1.23)	7.70 (-0.83)	7.76 (-0.67)	7.85 (-0.43)	7.89 (-0.33)	7.94 (-0.17)	7.95 (-0.17)	7.95 (-0.16)	7.95 (-0.17)
	4	7.56 (-1.33)	7.74 (-0.68)	7.74 (-0.72)	7.92 (-0.25)	7.99 (-0.05)	8.00 (-0.03)	7.99 (-0.06)	8.05 (0.10)	8.03 (0.05)
	10	7.05 (-2.76)	7.71 (-0.79)	7.90 (-0.30)	7.98 (-0.06)	8.02 (0.02)	7.97 (-0.11)	7.96 (-0.12)	7.99 (-0.06)	7.97 (-0.09)
	Inf	6.42 (-4.15)	7.69 (-0.78)	7.88 (-0.30)	7.98 (-0.07)	7.97 (-0.09)	7.92 (-0.19)	7.95 (-0.13)	7.89 (-0.27)	7.89 (-0.26)
25	0	7.58 (-1.03)	—	—	—	—	—	—	—	—
	0.5	7.51 (-1.32)	7.54 (-1.21)	7.48 (-1.35)	7.54 (-1.13)	7.55 (-1.07)	7.49 (-1.24)	7.48 (-1.27)	7.46 (-1.33)	7.46 (-1.34)
	1	7.71 (-0.62)	7.56 (-1.06)	7.64 (-0.80)	7.66 (-0.77)	7.71 (-0.62)	7.74 (-0.53)	7.73 (-0.57)	7.74 (-0.51)	7.75 (-0.49)
	2	7.41 (-1.46)	7.54 (-1.02)	7.59 (-0.86)	7.69 (-0.62)	7.72 (-0.52)	7.78 (-0.36)	7.78 (-0.36)	7.78 (-0.36)	7.78 (-0.37)
	4	7.37 (-1.62)	7.57 (-0.87)	7.57 (-0.92)	7.75 (-0.46)	7.82 (-0.26)	7.82 (-0.24)	7.81 (-0.28)	7.87 (-0.12)	7.85 (-0.17)
	10	6.82 (-3.16)	7.52 (-1.02)	7.71 (-0.54)	7.80 (-0.31)	7.83 (-0.24)	7.77 (-0.39)	7.76 (-0.40)	7.78 (-0.34)	7.77 (-0.38)
	Inf	6.09 (-4.75)	7.44 (-1.17)	7.65 (-0.62)	7.75 (-0.39)	7.74 (-0.39)	7.69 (-0.49)	7.71 (-0.44)	7.65 (-0.58)	7.65 (-0.57)

Notes: See Table 2.7 but for EL 20-based portfolios.

Table 2.10: Average returns of MZ R^2-based portfolios.

		δ								
c	η	0	0.6	0.8	0.9	0.94	0.98	0.99	0.999	1
0	0	8.43 (0.14)	—	—	—	—	—	—	—	—
	0.5	8.36 (-0.07)	8.37 (-0.04)	8.40 (0.04)	8.39 (0.01)	8.39 (0.03)	8.38 (-0.03)	8.39 (-0.00)	8.38 (-0.01)	8.39 (0.01)
	1	8.41 (0.06)	8.42 (0.12)	8.43 (0.13)	8.41 (0.09)	8.41 (0.08)	8.44 (0.18)	8.44 (0.17)	8.43 (0.15)	8.43 (0.15)
	2	8.51 (0.38)	8.45 (0.21)	8.47 (0.26)	8.47 (0.26)	8.45 (0.18)	8.45 (0.19)	8.50 (0.35)	8.48 (0.30)	8.49 (0.32)
	4	8.51 (0.37)	8.69 (0.94)	8.61 (0.72)	8.64 (0.74)	8.61 (0.66)	8.55 (0.47)	8.59 (0.60)	8.62 (0.69)	8.62 (0.70)
	10	8.65 (0.78)	8.55 (0.46)	8.49 (0.30)	8.49 (0.31)	8.57 (0.50)	8.58 (0.56)	8.61 (0.63)	8.62 (0.67)	8.63 (0.68)
	Inf	9.02 (1.39)	8.33 (-0.16)	8.24 (-0.41)	8.21 (-0.49)	8.34 (-0.11)	8.37 (-0.03)	8.40 (0.03)	8.49 (0.25)	8.54 (0.37)
5	0	8.26 (-0.10)	—	—	—	—	—	—	—	—
	0.5	8.19 (-0.30)	8.20 (-0.26)	8.23 (-0.18)	8.22 (-0.21)	8.23 (-0.20)	8.21 (-0.25)	8.22 (-0.22)	8.22 (-0.23)	8.22 (-0.22)
	1	8.24 (-0.16)	8.26 (-0.10)	8.26 (-0.08)	8.25 (-0.12)	8.25 (-0.13)	8.28 (-0.03)	8.28 (-0.05)	8.27 (-0.07)	8.27 (-0.07)
	2	8.34 (0.15)	8.29 (-0.01)	8.31 (0.06)	8.31 (0.06)	8.29 (-0.02)	8.29 (-0.01)	8.34 (0.15)	8.32 (0.09)	8.33 (0.11)
	4	8.33 (0.12)	8.52 (0.72)	8.45 (0.51)	8.48 (0.55)	8.46 (0.48)	8.39 (0.29)	8.43 (0.41)	8.46 (0.50)	8.46 (0.51)
	10	8.44 (0.45)	8.38 (0.25)	8.32 (0.08)	8.33 (0.10)	8.40 (0.30)	8.42 (0.37)	8.45 (0.45)	8.46 (0.49)	8.47 (0.50)
	Inf	8.70 (0.91)	8.07 (-0.66)	7.98 (-0.88)	7.94 (-0.98)	8.08 (-0.57)	8.12 (-0.42)	8.15 (-0.35)	8.24 (-0.12)	8.28 (-0.02)
10	0	8.09 (-0.33)	—	—	—	—	—	—	—	—
	0.5	8.02 (-0.53)	8.04 (-0.48)	8.07 (-0.40)	8.05 (-0.44)	8.06 (-0.42)	8.04 (-0.47)	8.05 (-0.44)	8.05 (-0.45)	8.05 (-0.44)
	1	8.07 (-0.38)	8.10 (-0.31)	8.10 (-0.29)	8.09 (-0.33)	8.08 (-0.35)	8.12 (-0.25)	8.11 (-0.26)	8.11 (-0.28)	8.11 (-0.28)
	2	8.17 (-0.08)	8.13 (-0.23)	8.15 (-0.15)	8.15 (-0.14)	8.12 (-0.22)	8.13 (-0.21)	8.18 (-0.05)	8.16 (-0.11)	8.17 (-0.10)
	4	8.15 (-0.14)	8.36 (0.51)	8.29 (0.30)	8.32 (0.35)	8.30 (0.29)	8.23 (0.10)	8.27 (0.22)	8.30 (0.31)	8.30 (0.32)
	10	8.24 (0.11)	8.21 (0.03)	8.15 (-0.13)	8.16 (-0.11)	8.24 (0.11)	8.26 (0.18)	8.29 (0.28)	8.31 (0.31)	8.31 (0.32)
	Inf	8.38 (0.41)	7.80 (-1.16)	7.71 (-1.35)	7.67 (-1.47)	7.81 (-1.04)	7.86 (-0.81)	7.89 (-0.73)	8.00 (-0.49)	8.03 (-0.40)
15	0	7.92 (-0.57)	—	—	—	—	—	—	—	—
	0.5	7.86 (-0.76)	7.87 (-0.69)	7.90 (-0.62)	7.89 (-0.66)	7.89 (-0.64)	7.88 (-0.69)	7.89 (-0.66)	7.88 (-0.67)	7.89 (-0.66)
	1	7.90 (-0.60)	7.93 (-0.52)	7.94 (-0.50)	7.92 (-0.54)	7.92 (-0.56)	7.95 (-0.46)	7.95 (-0.47)	7.94 (-0.49)	7.94 (-0.49)
	2	8.00 (-0.31)	7.97 (-0.44)	7.99 (-0.36)	7.99 (-0.34)	7.96 (-0.42)	7.96 (-0.42)	8.02 (-0.25)	8.00 (-0.32)	8.00 (-0.30)
	4	7.97 (-0.39)	8.20 (0.30)	8.13 (0.09)	8.16 (0.16)	8.14 (0.11)	8.07 (-0.08)	8.11 (0.03)	8.15 (0.13)	8.15 (0.13)
	10	8.03 (-0.22)	8.04 (-0.18)	7.98 (-0.34)	7.99 (-0.31)	8.07 (-0.08)	8.10 (-0.00)	8.14 (0.10)	8.15 (0.13)	8.16 (0.15)
	Inf	8.06 (-0.09)	7.53 (-1.66)	7.45 (-1.81)	7.40 (-1.96)	7.55 (-1.50)	7.60 (-1.19)	7.64 (-1.11)	7.75 (-0.87)	7.78 (-0.78)
20	0	7.75 (-0.80)	—	—	—	—	—	—	—	—
	0.5	7.69 (-0.98)	7.70 (-0.91)	7.73 (-0.84)	7.72 (-0.88)	7.73 (-0.86)	7.71 (-0.92)	7.72 (-0.88)	7.71 (-0.89)	7.72 (-0.88)
	1	7.74 (-0.82)	7.77 (-0.73)	7.77 (-0.71)	7.76 (-0.75)	7.75 (-0.77)	7.79 (-0.67)	7.78 (-0.68)	7.78 (-0.70)	7.78 (-0.70)
	2	7.83 (-0.53)	7.80 (-0.66)	7.82 (-0.56)	7.83 (-0.54)	7.80 (-0.62)	7.80 (-0.62)	7.86 (-0.45)	7.84 (-0.52)	7.84 (-0.51)
	4	7.79 (-0.64)	8.04 (0.09)	7.97 (-0.11)	8.00 (-0.03)	7.99 (-0.07)	7.92 (-0.26)	7.96 (-0.15)	7.99 (-0.06)	7.99 (-0.06)
	10	7.82 (-0.56)	7.86 (-0.40)	7.82 (-0.56)	7.83 (-0.52)	7.91 (-0.27)	7.94 (-0.18)	7.98 (-0.08)	7.99 (-0.04)	8.00 (-0.03)
	Inf	7.75 (-0.60)	7.27 (-2.15)	7.19 (-2.27)	7.13 (-2.45)	7.28 (-1.96)	7.34 (-1.57)	7.39 (-1.48)	7.50 (-1.24)	7.53 (-1.16)
25	0	7.58 (-1.03)	—	—	—	—	—	—	—	—
	0.5	7.52 (-1.21)	7.54 (-1.12)	7.56 (-1.05)	7.55 (-1.09)	7.56 (-1.08)	7.54 (-1.14)	7.55 (-1.10)	7.55 (-1.11)	7.55 (-1.09)
	1	7.57 (-1.04)	7.60 (-0.94)	7.61 (-0.91)	7.59 (-0.96)	7.59 (-0.98)	7.62 (-0.88)	7.62 (-0.89)	7.61 (-0.91)	7.61 (-0.91)
	2	7.66 (-0.76)	7.64 (-0.87)	7.66 (-0.77)	7.67 (-0.74)	7.64 (-0.82)	7.64 (-0.82)	7.70 (-0.65)	7.68 (-0.73)	7.68 (-0.71)
	4	7.61 (-0.90)	7.87 (-0.12)	7.81 (-0.32)	7.84 (-0.23)	7.83 (-0.25)	7.76 (-0.44)	7.80 (-0.33)	7.83 (-0.25)	7.83 (-0.25)
	10	7.61 (-0.91)	7.69 (-0.61)	7.65 (-0.77)	7.66 (-0.72)	7.75 (-0.46)	7.79 (-0.36)	7.82 (-0.25)	7.84 (-0.22)	7.84 (-0.20)
	Inf	7.43 (-1.13)	7.00 (-2.64)	6.92 (-2.72)	6.87 (-2.93)	7.02 (-2.41)	7.08 (-1.95)	7.14 (-1.86)	7.26 (-1.60)	7.28 (-1.54)

Notes: See Table 2.7 but for MZ R^2-based portfolios.

3 Measurement error sensitivity of loss functions for distribution forecasts

Abstract

We examine the sensitivity of loss functions—equivalently called scoring rules—for distribution forecasts in two dimensions: linear rescaling of the data and the influence of measurement error on the forecast evaluation outcome. First, we show that all commonly used scoring rules for distribution forecasts are robust to rescaling the data. Second, it is revealed that the forecast ranking based on the continuous ranked probability score is less sensitive to measurement error than the log score. Our theoretical results are complemented by a simulation study aligned with quarterly US GDP growth data and an empirical application forecasting realized variances of 28 Dow Jones Industrial Average constituents. In line with its proven gross-error insensitivity, the ranking of the continuous ranked probability score is the most consistent between evaluations based on the true outcome and the observations with measurement error.

3.1 Introduction

Distribution forecasts provide means to communicate the uncertainty that comes along predicting future outcomes as opposed to point predictions. However, in economic forecasting the true outcome is often a latent variable and, thus, predictions have to be evaluated against noisy proxy observations. A leading example in communicating the uncertainty of GDP estimates and forecasts is the Bank of England's Monetary Policy Committee who reports probabilistic forecasts of inflation rates and GDP since February 1996 and November 1997, respectively.[1] In this paper, we analyze the sensitivity of distribution forecast evaluation in settings in which the predictand is observed with measurement error or simply measured on different scales.

For assessing forecast accuracy, Gneiting and Raftery (2007) promote proper scoring rules—equivalently called loss functions—as an incentive for stating honest beliefs about future outcomes. We provide an overview of proper scoring rules for distribution forecasts in Table 3.1.

[1] https://www.bankofengland.co.uk/inflation-report/inflation-reports

The table includes both widely used scoring rules like the log score (LogS) and the continuous ranked probability score (CRPS) but also lesser known ones; for example, the power score or the pseudo spherical score. Correctly scaled versions of the latter two scores include the log score as a limiting case (Good, 1971). The quadratic score (QS) can be thought of being a continuous analogue of the widely-used Brier score for discrete variables (Brier, 1950). The weighted logarithmic scoring rule by Amisano and Giacomini (2007) is not listed as it is an improper scoring rule (Gneiting and Ranjan, 2011).

Table 3.1: Common proper scoring rules for distribution forecasts

Name	Definition	Shape	PDF	CDF
Log(arithmic) score	$\mathrm{LogS}(f,y) = -\log f(y)$	–	✓	–
Censored likelihood score	$\mathrm{CLS}(f,y) = -\mathbb{1}_{\{y\in A\}} \log f(y) - \mathbb{1}_{\{y\in A^C\}} \log\left(\int_{A^C} f(s)\ \mathrm{d}s\right)$	$A \subset \mathbb{R}$	✓	–
Power score	$\mathrm{PS}_\gamma(f,y) = -\gamma f(y)^{\gamma-1} + (\gamma-1)\|f\|_\gamma^\gamma$	$\gamma > 1$	✓	–
Quadratic score	$\mathrm{QS}(f,y) = \mathrm{PS}_2(f,y) = -2f(y) + \|f\|_2^2$	–	✓	–
Pseudo spherical score	$\mathrm{PseudoS}_\delta(f,y) = -\frac{f(y)^{\delta-1}}{\|f\|_\delta^{\delta-1}}$	$\delta > 1$	✓	–
Spherical score	$\mathrm{SphS}(f,y) = \mathrm{PseudoS}_2(f,y) = -\frac{f(y)}{\|f\|_2}$	–	✓	–
Hyvärinen score	$\mathrm{HyvS}(f,y) = 2\frac{f''(y)}{f(y)} - \left(\frac{f'(y)}{f(y)}\right)^2$	–	✓	–
Continuous ranked probability score	$\mathrm{CRPS}(F,y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}_{\{z>y\}}(z))^2\ \mathrm{d}z$	–	–	✓

Notes: The table provides an overview of proper scoring rules for distribution forecasts. To the best of our knowledge, the corresponding scoring rules were shown to incentivize stating honest beliefs about future outcomes by the following authors. LogS: Good (1952), CLS: Diks, Panchenko, and Dijk (2011), PS and QS: Buehler (1971), SphS: Buehler (1971), PseudoS: Good (1971), HyvS: Hyvärinen (2005), CRPS: Matheson and Winkler (1976). The latter two columns indicate whether the scoring rules are based on the probability density function or the cumulative distribution function. $\|\cdot\|_\gamma$ refers to the L^γ norm.

We show that all scoring rules in Table 3.1 are robust to a linear rescaling of the data. This is achieved by introducing the notion of scaling-invariance for loss functions, which is a slightly more general definition of homogeneity than the one used in Patton (2011) for point forecasts. However, this result cannot be generalized to hold for every proper loss function for distribution forecasts, as it is possible to construct proper scoring rules that are not robust to rescaling by combining scoring rules of different degrees of scaling-invariance.

When evaluating forecasts using scoring rules, a major problem is that the observed value of the target variable is not necessarily equal to the true predictand's value in many economic situations. For example, in case of additive measurement error, the variance of the observations is always higher than the variance of the true predictand. This causes proper scoring rules to prefer distribution forecasts with larger variances than the conditional variance of the predictand of interest. In order to address this misalignment, Ferro (2017) and Naveau and Bessac (2018) propose to calculate error-corrected scoring rules. Their underlying idea for eliminating the bias introduced due to measurement error is to use the difference between the expected loss when employing the noisy proxy and the expected loss when employing the true predictand. However, this approach is tied to knowing the true predictand's distribution, the underlying error distribution, and the specific forecast distribution at hand. Such restrictive assumptions are typically not fulfilled in economic applications. In case of the additive noise model, we only have strictly proper error-corrected scoring rules if every entity follows a Gaussian distribution.

Therefore, we choose a different approach and quantify the expected deviation in loss due to employing a noisy proxy for forecast evaluation. A desirable property of loss functions with respect to our measure of loss-induced bias, the expected absolute deviation function, is that measurement errors should not be able to increase the expected loss beyond any boundary. Loss functions that have this property will be called gross-error insensitive. Our measure is linked to the theory of robust estimators in the notion of Hampel (1968, 1971) and, more specifically, the influence function of an estimator. The influence function of an estimator quantifies the change of an estimate due to an infinitesimal distortion in the observations. However, we are interested in the change of expected forecast rankings and, thus, employ an expected value framework instead of an infinitesimal approach. The quadratic score and the CRPS turn out to be gross-error insensitive but the log score is not.

Our results can be linked to the literature on robust estimation in regard to employing proper scoring rules for M-estimation (Dawid, Musio, and Ventura, 2016). Basu et al. (1998) show that power scoring rules that deviate from the limiting case of the log score are more robust to small outliers in the data. On the contrary, Kanamori and Fujisawa (2015) find that large outliers are best coped with using the pseudo spherical score. A good overview of this branch of the literature can be found in Ovcharov (2017).

Our theoretical results are illustrated by a simulation study and an empirical application. We use these to bridge the gap from observing losses in absolute terms to testing for equal predictive ability in the form of Diebold-Mariano (DM) or Giacomini-White (GW) tests (Diebold and Mariano, 1995; Giacomini and White, 2006). For simplicity, we will refer to tests for equal forecast performance always as DM tests as the paper focuses on comparing forecasts, not models. However, the test outcomes in the simulation and empirical application could also be understood to be GW tests as they fulfill the criteria of a fixed/rolling window estimation scheme.

With a data-generating process aligned with real-time United States (US) gross domestic product (GDP) growth data, we provide simulation-based evidence that using the CRPS in DM tests is less sensitive to observational error than using the log score. In the simulation, the employed measurement error process mimics the empirically observed level of measurement error present in second-release data of GDP.[2] We see that frequent but possibly small level of observational error causes the quadratic score to perform even worse though it is a gross-error insensitive loss function. The outcome of the quadratic score being "only" gross-error insensitive but not "small-error insensitive" is confirmed in the empirical application.

In the empirical application, we evaluate distribution forecasts for asset price volatility of 28 Dow Jones Industrial Average (DJIA) constituents that are evaluated against two different volatility proxies. As in the simulation, the DM test statistics display the CRPS to be the least sensitive across different outcome estimates.

The outline of this paper is as follows. Section 3.2 introduces our theoretical results and Section 3.3 validates their implications in a simulation study. Section 3.4 presents our empirical application, and is followed by a discussion in Section 3.5. All proofs are deferred to Appendix 3.6.1. We restrict the analysis to real-valued random variables and use the terms loss function and scoring rule interchangeably while lower values are always considered to be associated with more precise forecasts. Distribution forecasts in form of densities are always denoted by lower case letters. Distribution forecasts in form of distribution functions are denoted by upper case letters.

3.2 Theory

In this paper, we consider loss functions for distribution forecasts that are proper, meaning that a forecaster's expected loss is minimized if she states the true conditional distribution of the outcome variable (Gneiting and Raftery, 2007). As a consequence, proper loss functions incentivize stating honest beliefs about future outcomes which is evidently a desirable property of a good forecast evaluation criterion. In Table 3.1, it can be seen that there is a wide range of different proper scoring rules for distribution forecasts: On the one hand, the log score's and

[2] In our context of unconditional forecast evaluation, we subsume every deviation from the latent true predictand as measurement error in contrast to discussions whether macroeconomic revisions are "news" or "noise" (e.g., Faust, Rogers, and Wright, 2005; Aruoba, 2008). Similarly, Clements and Galvao (2018) employ data revision uncertainty in macroeconomic models for improving density forecasts. We will restrict ourselves to forecast evaluation.

the Hyvärinen score's realized loss only depends on the value of the density function and its derivatives at the outcome. On the other hand, the power score and the pseudo spherical score employ the likelihood and the L^γ norm as an additional measure of sharpness that is independent of the outcome. Contrary to all other loss functions listed in Table 3.1, the CRPS can also be calculated without knowing the density function which is especially helpful when dealing with Markov Chain Monte Carlo output. In the following, we examine the influence of rescaling the data and measurement error on forecast evaluation.

3.2.1 Rescaling the data and forecast rankings

In economic applications, the units of measurement may vary; for example, returns are either reported in decimals or percentages. Therefore, it is of interest if a simple linear rescaling of the data may change the ranking of forecasts and for which loss functions the ranking is robust to such data transformations. We will see that not every proper scoring rule has this property.

Regarding point forecasts, Patton (2011) shows that the expected ranking for homogeneous loss functions does not change due to rescaling. In the context of evaluating a point forecast $\hat{y}$ for realization y, the homogeneity of a loss function L of degree k is given by

$$L(\lambda\hat{y}, \lambda y) = \lambda^k L(\hat{y}, y) \text{ for all } \lambda > 0.$$

As we consider distribution forecasts, we introduce a more general notion of the term homogeneity.

Definition 3.1. *A scoring rule L for a distribution forecast F is said to be scaling-invariant of order $k \in \mathbb{R}$ if for all $\lambda > 0$ and all possible realizations y we have that*

$$L(F_\lambda, \lambda y) = \lambda^k L(F, y) + C(\lambda, y)$$

with $F_\lambda(y) = F(y/\lambda)$ and C being a function of λ and y.

Definition 3.1 says that if the distribution forecasts and the observations are scaled up or down by the same constant $\lambda > 0$, the loss may increase or decrease multiplicatively in λ and a shift $C(\lambda, y)$ that does not depend on the forecast. Instead, the shift should only depend on the realization and the scaling factor λ. As scaling-invariance is not merely defined for density but for distribution forecasts in general, the notion of homogeneity of loss functions discussed in Patton (2011) is nested if one considers distribution forecasts with point mass one in the outcome space:

Remark 3.1. *If we consider a point forecast y_0, $F(x) = \mathbb{1}_{\{y_0 \leq x\}}(x)$, Definition 3.1 reduces to the homogeneity of loss functions considered in Patton (2011, Proposition 3) with $C \equiv 0$.*

The result that all commonly used loss functions in Table 3.1 are scaling-invariant is stated below in Proposition 3.1.

Proposition 3.1. *The log score, censored likelihood score, power score, pseudo spherical score, Hyvärinen score, and continuous ranked probability score are scaling-invariant of the following order. LogS: zero, CLS : zero,* PS_γ*:* $1-\gamma$*,* $\mathrm{PseudoS}_\delta$*:* $(1-\delta)/\delta$*,* HyvS*: two,* CRPS*: one.*

The following Proposition 3.2 is an analogous result to the findings laid out in Patton (2011, Proposition 3). Instead of focusing on homogeneous loss functions for point forecasts, we now examine these for distribution forecasts.

Proposition 3.2. *(a) The ranking of two distribution forecasts by expected loss is invariant to a rescaling of the data if the loss function is scaling-invariant.*

(b) The ranking of two distribution forecasts by expected loss may not be invariant to a rescaling of the data if the loss function is not scaling-invariant.

Hence, it is shown that all commonly used loss functions for distribution forecasts are robust to rescaling the data because they fulfill the notion of scaling-invariance introduced in Definition 3.1. The proof of the second part of Proposition 3.2 employs the sum of the LogS and the CRPS as a proper scoring rule. As these two loss functions are scaling-invariant of different order, the ranking can be reversed by rescaling the data even in the simple case of Gaussian distribution forecasts.

It is worthwhile to note that the definition of scaling-invariance implies changes in loss differences to be only scaled up or down by λ^k and, hence, the test outcome for equal predictive ability by means of a DM or GW test statistic will be unaffected by the rescaling. A similar notion has been discussed in Patton, Ziegel, and Chen (2019) where they consider scoring rules for value-at-risk and expected shortfall that generate homogeneous loss differences of order zero.

3.2.2 Measurement error and forecast rankings

After assessing robustness with respect to rescaling the data, we turn to assessing the impact of additive measurement error present in many economic contexts. Our aim is to examine for which loss functions the ranking of competing forecasts is less likely to reverse in the presence of observational error and not to propose altered scoring rules that compensate the noise-induced bias. In this section, we show theoretically that the rankings of forecasts by the CRPS and, to a lesser extent, the quadratic score are less sensitive to measurement error than the log score. We restrict ourselves to these three leading examples because the log score and CRPS are the most widely-employed scoring rules for distribution forecasts and the quadratic score is interesting due to its relationship to the log score.[3] The implications will be analyzed via a simulation study in Section 3.3 and an empirical application in Section 3.4.

In the following, the setup is always that there is a true latent predictand random variable Y and corresponding observations $\widetilde{Y}$ with measurement error, $\widetilde{Y} = Y + U$. We want to identify forecasts that are superior in forecasting Y but we can only use $\widetilde{Y}$ for forecast evaluation.[4] For quantifying the induced bias due to noise, we define the *expected absolute deviation function.*

Definition 3.2. *The expected absolute deviation function of a loss function L with respect to the distribution forecast F and forecast error distribution G at realization y is defined as*

$$\mathrm{EADF}(L, F, y, G) = \mathbf{E}_U \left| L(F, y + U) - L(F, y) \right|,$$

where U is G-distributed, $U \sim G$.

The definition of the expected absolute deviation function is inspired by the notion of the influence function in robust statistics (Hampel et al., 1986, p. 84). The influence function has an intuitive interpretation as describing the effect of infinitesimal observational errors on the asymptotic value of an estimator. Likewise, the expected absolute deviation function quantifies the bias in the expected loss implied by a certain error distribution G with respect to a certain distribution forecast F. The main difference is that in our definition we are considering expected

[3] In this paper, we only use analytical expression for the quadratic score of the normal and the log-normal distribution but Appendix 3.6.2 includes additional results for the mixture of normals, student-t, generalized beta, and two-piece normal distribution because they are widely applied in macroeconomics and financial econometrics. For all distributions (besides the mixture of normals distribution) we report the more general power score by calculating the densities' L^γ norm instead of the L^2 norm.

[4] We restrict ourselves to the additive measurement error specification because to us it appears to be the most prevalent scenario in economics. Future research could extend our results to a multiplicative error structure.

deviations with respect to a certain error distribution. Even though it is not the focus of this paper, we will discuss the extension of our results in terms of finite-sample breakpoints (Hampel, 1968, 1971; Huber, 1984) in Remark 3.2 at the end of this section.

Definition 3.2 is bounded to one distribution forecast F and one forecast error distribution G. However, in most forecasting situations one is confronted with time-varying distribution forecasts and sometimes even time-varying measurement error. In order to address this problem, we define a summary statistic of the expected absolute deviation function with respect to classes of forecast and noise distributions. For example, when forecasting macroeconomic variables using Bayesian vector autoregressive models, forecasts may have time-varying parameters but are part of a certain family of distributions; for example, the class of mixture of normal distributions. Similarly, one may not be able to model the measurement error accurately but may be sure that the distribution of the measurement has at least a finite second moment. For assessing the overall impact of measurement errors in this scenario, we want to quantify the expected absolute deviation with respect to all possible distribution forecasts (e.g., the class of mixture of normal distributions) and all possible error distributions that we assume to be realistic (e.g., the class of all distributions with finite second moment). This motivates our next definition of loss function sensitivity.

Definition 3.3. *The gross-error sensitivity γ^* of a loss function L with respect to a forecast distribution F in $\mathcal{P}_F$ and an error distribution G in $\mathcal{P}_G$ is defined as*

$$\gamma^* = \sup_y \mathrm{EADF}(L, F, y, G).$$

Moreover, we call the loss function L gross-error insensitive with respect to forecast distributions F in $\mathcal{P}_F$ and error distributions G in $\mathcal{P}_G$ if $\gamma^* < \infty$ for all $F \in \mathcal{P}_F$ and $G \in \mathcal{P}_G$.

The gross-error sensitivity γ^* can be interpreted as an upper bound to the worst expected absolute deviation of the loss function from its "true" value without observational error.

As it was the case in the definition of the expected absolute deviation function, our notion of gross-error sensitivity is closely related to the (infinitesimal) gross-error sensitivity considered by Hampel et al. (1986, p. 87).

The implications for the forecast ordering can be seen from the following example. Let F_1 and F_2 denote two different forecasts and $\widetilde{Y} = Y + U$ the sum of the true predictand random variable Y and a measurement error $U \sim G$. Furthermore, assume that F_1 is the better forecast in expectation; that is, there exists $\overline{L} > 0$ such that

$$\mathbf{E}_Y[L(F_2, Y)] - \mathbf{E}_Y[L(F_1, Y)] = \overline{L} > 0.$$

Given that the EADF is gross-error insensitive, we can find a sufficient condition for obtaining the same ranking using the noisy proxy; that is,

$$\mathbf{E}_{\widetilde{Y}}[L(F_2, \widetilde{Y})] - \mathbf{E}_{\widetilde{Y}}[L(F_1, \widetilde{Y})] > 0. \tag{3.1}$$

We assume L to be gross-error insensitive with respect to our forecasts F_1, F_2 and error distribution G (i.e., $\gamma^*_{F_1}$ for F_1, $\gamma^*_{F_2}$ for F_2). If we further assume that $\max\{\gamma^*_{F_1}, \gamma^*_{F_2}\} < \overline{L}/2$, then

$$\mathbf{E}_{Y,U}\left[\left|L(F_2, \widetilde{Y}) - L(F_2, Y)\right| + \left|L(F_1, \widetilde{Y}) - L(F_1, Y)\right|\right] \leq 2\gamma^* < \overline{L}. \tag{3.2}$$

Hence, if $\gamma^*_{F_1}$ and $\gamma^*_{F_2}$ are small enough relative to the forecast loss difference $\overline{L}$, the expected ranking is ensured to stay the same as Equation (3.2) implies Equation (3.1) to hold.

Our approach of discussing the bias in forecast evaluation introduced by measurement error is different from calculating error-corrected scoring rules as in Ferro (2017) and Naveau and Bessac (2018) that are only available for a handful of concrete pairs of forecast and error distributions. We highlight this difference by looking at the leading example of error-corrected scoring rules in Ferro (2017): Assume that the forecast distribution F is a normal distribution with mean μ and variance σ^2. Additionally, the true predictand Y is normally distributed with mean μ_0 and variance σ_0^2 and the independent additive measurement error U is a zero-mean normally distributed random variable with variance c^2. Then, the expected score of F with respect to $\widetilde{Y} = Y + U$ is given by

$$\mathbf{E}_{\widetilde{Y}}[\mathrm{LogS}(F, \widetilde{Y})] = \frac{1}{2}\log(2\pi) + \log\sigma + \frac{(\mu - \mu_0)^2 + \sigma_0^2 + c^2}{2\sigma^2} \tag{3.3}$$

which exceeds the expected score of F with respect to Y,

$$\mathbf{E}_Y[\mathrm{LogS}(F, Y)] = \frac{1}{2}\log(2\pi) + \log\sigma + \frac{(\mu - \mu_0)^2 + \sigma_0^2}{2\sigma^2}, \tag{3.4}$$

by an amount of $c^2/(2\sigma^2)$. Whereas Equation (3.4) is minimized for $\mu = \mu_0$ and $\sigma^2 = \sigma_0^2$, Equation (3.3) is minimized if $\mu = \mu_0$ but $\sigma^2 = \sigma_0^2 + c^2$. This is an example that the log score as a proper scoring rule favors a model that predicts $\widetilde{Y}$ (the noisy variable) instead of a model that predicts Y (the noise-free variable). Given our knowledge about the difference in $\mathbf{E}_{\widetilde{Y}}[\mathrm{LogS}(F,\widetilde{Y})]$ and $\mathbf{E}_Y[\mathrm{LogS}(F,Y)]$, Ferro (2017) defines the error-corrected log score $\mathrm{LogS}_c(F,Y)$ in this scenario to be

$$\mathrm{LogS}_c(F,y) = \mathrm{LogS}(F,y) - \frac{c^2}{2\sigma^2}. \tag{3.5}$$

By construction, the expected core in Equation (3.5) is again minimized if $\mu = \mu_0$ and $\sigma^2 = \sigma_0^2$. Using this error-corrected scoring rule in contrast to the "vanilla" log score, one incentivizes stating honest beliefs about the true predictand Y instead of $\widetilde{Y}$. However, this comes at a cost. Evidently, Equation (3.5) readily implies that in empirical applications c has to be correctly specified in order to alter the incentives into the right direction. Moreover, error-corrected scoring rules in our additive observation error model do not need to be as easily derived as suggested by Equation (3.5). Ferro (2017) and Naveau and Bessac (2018) derive two error-corrected strictly proper scoring rules for additive observational error: the error-corrected log score and CRPS—but only so for Gaussian distribution forecasts with independently and normally distributed noise. Alternatively, Ferro (2017) proposes an error-corrected Dawid-Sebastiani score (Dawid and Sebastiani, 1999) under slightly more general terms but this score does not discriminate between forecast distributions with the same first and second moments. However, this means that the evaluation of rivaling distribution forecasts is reduced to a joint point forecast evaluation of predictive mean and variances.

Now, we come to the main proposition of this paper regarding the gross-error sensitivity of selected scoring rules in Table 3.1.

Proposition 3.3. *(a) The log score is generally not gross-error insensitive for any class of forecasts that includes the class of normal distributions and error distributions with finite second moment.*

(b) The quadratic score is gross-error insensitive with respect to the subclass of all forecasts $f \in L^2$ which fulfill $f(x) \leq \bar{f}$ for some individual upper bound $\bar{f} > 0$ and arbitrary error distributions G.

(c) The CRPS is gross-error insensitive with respect to forecast distributions F that have finite first moment and error distributions G that have finite first absolute moment.

The counterexample in the proof of part (a) in Proposition 3.3 is a very general one for a

Gaussian distribution forecast and a general error distribution U for which the only assumption is $\mathbf{E}[U^2] < \infty$.[5]

The proofs of part (b) and (c) in Proposition 3.3 provide additional insight: We show that the upper bound of the CRPS with respect to error distributions U with finite first moments is independent of the forecast distribution itself,

$$\mathrm{EADF}(\mathrm{CRPS}, F, y, G) \leq \mathbf{E}_U \left| U \right| .$$

This is in contrast to the upper bound of the EADF of the QS,

$$\mathrm{EADF}(\mathrm{QS}, f, y, g) \leq 2\overline{f}$$

that depends on the upper bound of the density forecast f. This implies that for different noise-to-signal scenarios either the QS or CRPS will perform better. In our simulations and the empirical section, the CRPS reigns supreme.

In comparison to Patton (2011), we have not derived a class of loss functions for which the ranking of two models will be preserved in expectation when using a noisy proxy for forecast evaluation. However, we can say that the ranking is ensured to be less sensitive to measurement error for some loss functions than for others.

An insight on the negative result regarding the log score can be gained from the literature on M-estimation. The maximum likelihood estimator achieves the Cramér-Rao lower bound and can be interpreted to put equal weight on each observation (Basu et al., 1998, p. 551)—even distorted observations. For deriving their results, Basu et al. (1998) reinterpret maximum likelihood estimation in terms of minimizing the Kullback-Leibler divergence which itself is the divergence associated with the log score. Given that the log score is the limiting case of the appropriately scaled power score for $\gamma \to 1$, they examine an efficiency/robustness trade-off. Another interpretation of our results is that if one wants to evaluate distribution forecasts in the presence of noise and is interested in identifying good forecasts for the true predictand, one should use a loss function that is not as discriminatory between distributions as the log score is.

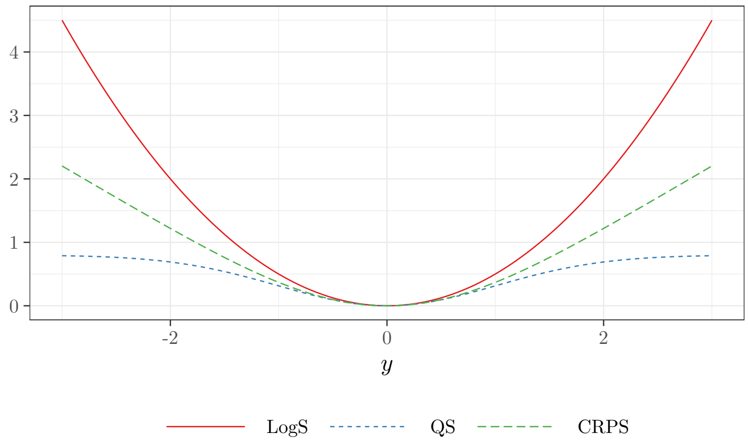
Remark 3.2. *Beyond the infinitesimal approach of error sensitivity that was the blue print for our definition of expected loss deviation, Hampel (1968, 1971) and Huber (1984) also discuss*

[5] Note that a similar argument as in the proof of part (b) could be made to prove the gross-error insensitivity of the pseudo spherical score.

finite sample characteristics of estimators: Given a finite sample $x_1, \ldots, x_n$, and the corresponding sample mean and median it can be shown that even distorting a single observational value may cause the sample mean to increase by an arbitrarily large number whereas for the sample median one would need to change $n/2$ observations to observe such a distortion. Hence, the sample mean is said to have a breakdown point of zero whereas the sample median has a breakdown point of 1/2 and this notion can be extended to finite sample losses.

For comparing the influence of changing a single observation on loss differences, Figure 3.1 shows that the log score and the CRPS may increase above any upper bound if the realization is far enough in the tails of the predictive distribution. On the contrary, the maximum value of the quadratic score is always the squared L^2 norm of the predictive density. In this sense, the log score and the CRPS have a "loss breakdown point" of zero whereas for the quadratic score the corresponding entity would be infinite.

Figure 3.1: LogS(f, y), QS(f, y) and CRPS(f, y) in case of a standard Gaussian density forecast f.



Note: Losses are standardized by subtracting their values at $y = 0$.

Summing up the theoretical section, we proved the robustness of all commonly used scoring rules for distribution forecasts to a linear rescaling of the data and the gross-error insensitivity of the quadratic score and CRPS in comparison to the log score. The results hold even for biased estimates of the true predictand and under mild regularity conditions on the additive error process. However, in comparison to the results of Patton (2011) regarding point forecasts, our findings do not imply that the expected forecast ranking using the noisy proxy always equals the forecast ranking using the true latent outcome. Instead, it is more likely to coincide for the QS and CRPS than for the log score.

3.3 Simulation

We evaluate our theoretical results in a simulation study tailored to US GDP growth data, which is the most prominent example for a time series featuring revision cycles in macroeconomics (Croushore, 2011). In the simulation, we examine the influence of different degrees of measurement error on the test outcomes of forecast rankings. The result is that the gross-error insensitivity of the CRPS leads to more stable DM test statistics in the presence of observational error.

3.3.1 ARMA model and observational error

The general idea is that we simulate a true underlying process which we want to forecast but, as often in practice, we only observe the outcome measured with error. In the simulation, we want to compare the alignment of forecast rankings with respect to three different measurement error scenarios: small and large continuously added noise and infrequent gross errors. The data-generating process is an ARMA(1,1),

$$Y_t = 0.7Y_{t-1} - 0.38\varepsilon_{t-1} + \varepsilon_t, \tag{3.6}$$

with $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon^2 = 4.43$. The parameters are chosen by maximum likelihood estimation on final quarterly US GDP growth rates in between 1989Q1 to 2015Q4.[6] In each of the 2,000 simulation runs, we simulate $T = P + R$ observations with $P = R = 100$. This corresponds to a scenario with each 25 years of quarterly data for estimation and forecast evaluation.

The observations used for model estimation and forecast evaluation are contaminated observations $\widetilde{Y}_t = Y_t + U_t$. In Scenario 1, we set $U_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_1^2)$. We choose $\sigma_1 = 1.38$, the empirical standard deviation of the observed difference between the second release and the 12th release of GDP growth estimates. Hence, this scenario corresponds to the case of the second release measurement error of GDP growth for which the null hypothesis of normality using the Shapiro-Wilk test is generally not rejected with a p-value of 0.94. In Scenario 2, we choose to simulate an even larger Gaussian measurement error with $\sigma_2 = 2\sigma_1$ representing the case of a very low signal-to-noise ratio. The measurement error in Scenario 3 is a normally distributed random variable with a variance larger than in Scenario 1 and 2, $\sigma_3^2 > \sigma_1^2$, $\sigma_3 = 4$, times an independent Bernoulli random variable for which the success probability is chosen such that $\mathbf{Var}(U_t) = \sigma_1^2$

[6] The data used for calibration is explained in more detail in Appendix 3.6.3.

in both Scenario 1 and 3,

$$U_t \sim \mathcal{N}(0, \sigma_3^2) \times Bernoulli(1 - \sqrt{1 - \sigma_1^2/\sigma_3^2}).$$

This can be interpreted as a scenario with rare but possibly large measurement errors.

On the simulated data, we compare two forecasting models. First, an ARMA(1,1) model is fitted on the first 100 "quarters" of the contaminated observations $\widetilde{Y}_t$ in each simulation run. The conditional distribution forecast is thus given by a Gaussian density in which the mean is given by the conditional mean forecast of the fitted ARMA model and the conditional standard deviation is given by the maximum likelihood estimate for σ_ε. The second model is a simple time-invariant Gaussian forecast density for which the mean and variance are given by the sample mean and sample variance of $\widetilde{Y}_t$ in the estimation period. Thus, the expected ranking of proper scoring rules in the absence of measurement error favors the ARMA model as it is similar to the true underlying process (subject to estimation error) while the alternative model is severely misspecified.[7]

In each simulation run, we fit the two models on the contaminated observations $\widetilde{Y}_t$ and compare the forecast ranking outcomes with respect to $\widetilde{Y}_t$ and the true but assumed to be unobservable Y_t. At time t in simulation run j, we calculate the corresponding loss differences $d_{j,t}$ and $\widetilde{d}_{j,t}$ where values smaller than zero indicate that the ARMA forecasts are better. The log scores of our models are given by the logarithm of the normal density function evaluated at the outcome values. For calculating the quadratic score, we need the squared L^2-norm of the predictive normal distribution which is given by

$$\|f_{Norm(\mu,\sigma^2)}\|_2^2 = \frac{1}{2\sqrt{\pi}\sigma}.$$

For details see Appendix 3.6.2. The closed-form solution of the CRPS of a normal distribution forecast has been derived in Gneiting et al. (2005). Thereafter, we calculate the DM test statistics using Newey-West standard errors (Newey and West, 1987) in each simulation run j; that is, for both the loss differences $d_{j,t}$ using the true outcome and $\widetilde{d_{j,t}}$ using the proxy observations with

[7] In finite samples, severely misspecified models may perform better in forecast performance than correctly specified models. The setup is chosen such that this is avoided in our simulation.

measurement error, we have

$$t_j = \frac{\overline{d_{j,t}}}{\widehat{\mathbf{Var}(d_{j,t})}} \text{ and } \widetilde{t_j} = \frac{\overline{\widetilde{d_{j,t}}}}{\widehat{\mathbf{Var}(\widetilde{d_{j,t}})}}.$$

3.3.2 Simulation results

In Table 3.2, we report summary figures across the simulation runs. In the first column, the share of equal forecast rankings across all simulation runs j is given. The ranking of the models appears to be most stable for the CRPS. It has the largest share of equal rankings in Scenario 1 and 2. In Scenario 3 the CRPS is just barely trailing behind the QS. The results reported in the second column show that $\overline{t_j}$, the cross-simulational average of t_j, is the smallest for the log score. This is in line with the notion of the log score obtaining the highest power to distinguish among densities. In the third column we report the difference between the averages $\overline{t_j}$ and $\overline{\widetilde{t_j}}$. The bias in the test statistic due to employing $\widetilde{Y}_t$ instead of the latent Y_t is always the smallest for the CRPS. It is noteworthy that the log score performs the worst in Scenario 3, the scenario with infrequent gross errors. Here, the difference $\overline{\widetilde{t_j}} - \overline{t_j}$ is more than three times larger than the difference for the CRPS.

Third, we report the standard deviation of the differences $\widetilde{t_j} - t_j$. In Scenario 1 and 2, this measure of variability is the lowest for the CRPS. In Scenario 3, it is only the second lowest for the CRPS but closer in value to the QS than the LogS for which we observe the largest standard deviation of t-statistic differences. All in all, the test statistics based on the CRPS are the least affected by measurement error.

3.4 Empirical application

The empirical application in this paper is targeted at comparing two rivaling distribution forecasts of realized variances. In line with the simulation results in Section 3.3, the CRPS gives the most consistent results between the outcome measurement that is considered to be the “true” realization and another one with larger measurement error.

3.4.1 Forecasting volatility

We examine to what extent the ordering of two different models may change when using different volatility proxies for forecast evaluation. It is known that the precision of realized variance

Table 3.2: Out-of-sample forecast comparison for simulation study.

	Equal loss	DM test stat.		
	ranking	$\overline{t_j}$	$\overline{\widetilde{t_j}} - \overline{t_j}$	sd$(\widetilde{t_j} - t_j)$
Scenario 1: Normally distributed noise, $\sigma_1 = 1.38$				
LogS	0.93	-2.45	-0.72	0.61
QS	0.91	-2.00	-0.61	0.71
CRPS	**0.95**	-2.05	**-0.36**	**0.56**
Scenario 2: Normally distributed noise, $\sigma_2 = 2\sigma_1$				
LogS	0.81	-2.17	-1.51	1.23
QS	0.79	-1.64	-1.64	1.24
CRPS	**0.85**	-1.30	**-0.67**	**1.04**
Scenario 3: Infrequent outliers				
LogS	0.95	-2.36	-0.40	0.55
QS	**0.98**	-1.95	-0.12	**0.30**
CRPS	0.98	-2.08	**-0.11**	0.31

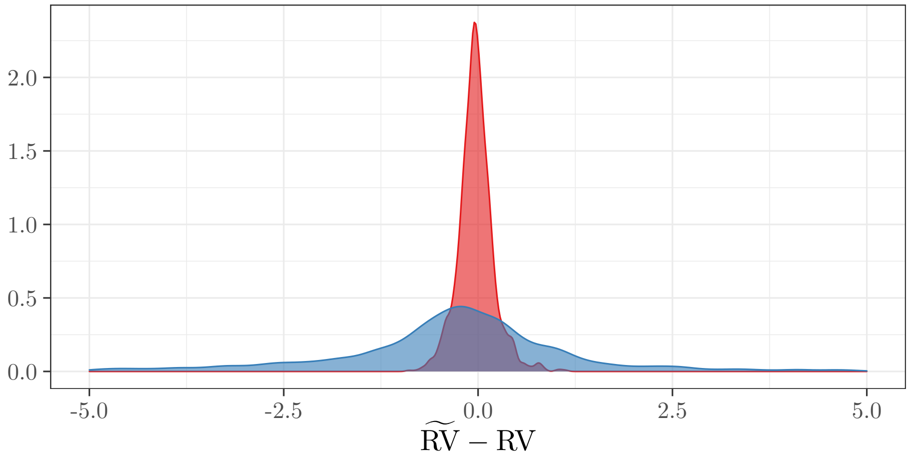
Notes: Scenario 1, 2, and 3 correspond to three different measurement error scenarios. The true data-generating process is an ARMA(1,1) calibrated on quarterly US GDP growth rates, see Equation (3.6). In Scenario 1, the observation error is a mean zero Gaussian random variable with $\sigma_1 = 1.38$. In Scenario 2, the measurement error's standard deviation is increased to $\sigma_2 = 2\sigma_1$. The third panel corresponds to the case of a mean zero normally distributed random variable with $\sigma_3 = 4$ times an independent Bernoulli random variable such that the variance of their product is equal to the measurement error variance in Scenario 1. Figures in bold indicate the largest share of equal rankings, the lowest observed average bias in the DM test statistic based on Newey-West standard errors, and the lowest standard deviation of the observed deviations in the test statistic. In total there are 2,000 simulation runs.

estimators decreases on days on which the integrated quarticity of the price process is higher (Barndorff-Nielsen and Shephard, 2002).

However, there are additional dimensions that make it important to have robust loss functions for volatility forecasting. The underlying data may be of different quality across time even for a single stock. For example, stocks that enter the S&P 500 are automatically traded more often as they become part of numerous ETFs along this step. Similarly, the quality of the volatility proxy may differ across stocks that enter a cross-sectional forecast evaluation. Last, given that high-frequency data is vast and has to be thoroughly cleaned another source of error is the cleaning process itself. Other authors discussing density forecasts for stock market volatility are Corsi et al. (2008), Corradi, Distaso, and Swanson (2009, 2011), Nonejad (2017), and Catania and Proietti (2020).

Our data set comprises 28 DJIA stocks in between 2000 and 2017. Based on intraday data, we compute at each day t and for each stock i two variance proxies denoted by $RV_{i,t}$ and $\widetilde{RV}_{i,t}$ that are based on either squared 5-minute or squared 15-minute returns. Additionally, we compute 15-minute semivariances, $\widetilde{RV}^{+}_{i,t}$ and $\widetilde{RV}^{-}_{i,t}$. More details on the data set can be found in Appendix 3.6.4. The average absolute difference between the two volatility proxies RV and $\widetilde{\text{RV}}$ in our sample across stocks and time is 0.65 when returns are measured in percentages relative to an average RV value of 2.73. As an example, we present smoothed densities of the differences in Figure 3.2 for the case of Apple Inc. The average level of RV for this stock is 5.04. We divide the sample into two subgroups: differences recorded on days on which the 5-minute realized variance is below (red) or above (blue) the empirical median inside the sample. It is evident that most of the differences in measurement are concentrated around zero. However, the dispersion of measurement errors is a lot higher in the high-volatility subsample depicted in blue.

Figure 3.2: Difference between volatility proxies for Apple Inc.



Notes: Smoothed histograms of the difference between the realized variance sampled on 15- or 5-minute returns; that is, $\widetilde{\text{RV}}$ - RV in case of Apple Inc. The density of observations below the empirical sample median is depicted in red, above the median in blue. The histogram is truncated discarding 219 observations. The time period is 2000–2017.

.

3.4.2 Volatility models

We employ two different heterogeneous autoregression (HAR) models for obtaining rivaling distribution forecasts of future volatility. In contrast to wide strands of the literature on volatility forecasting, we choose to model the realized variance process not in levels but in logs. The benefit of modeling the logarithm of RV is the empirical observation that the logarithm of RV is approximately normally distributed which enables us to compute closed-form distribution forecasts under the log-normality assumption.

First, our benchmark HAR model for each stock i is a simple autoregressive process that models tomorrow's log-volatility as a linear combination of past aggregated realized variances on a daily frequency:

$$\log \widetilde{RV}_{i,t+1} = \alpha_0 + \alpha_d \log \widetilde{RV}_{i,t} + \alpha_w \log \widetilde{RV}_{i,t-4:t} + \alpha_m \log \widetilde{RV}_{i,t-21:t} + \xi_{i,t},$$

in which the k-period cumulative realized variance is defined as $\widetilde{RV}_{i,t-k:t} = 1/k \sum_{j=0}^{k-1} \widetilde{RV}_{i,t-j}$ and $\mathbf{E}[\xi_{i,t}|\mathcal{F}_{t-1}] = 0$ with $\mathcal{F}_{t-1}$ denoting the information set up to time $t-1$. This model was introduced by Corsi (2009) and it readily implies one-step-ahead forecasts $\hat{\mu}^{HAR}_{i,t+1|t}$ for the log-mean of future $\widetilde{\mathrm{RV}}$. The one-step-ahead log-standard deviations $\hat{\sigma}^{HAR}_{i,t+1|t}$ of our predictive densities are chosen to be the empirical standard deviations of the residuals in the rolling estimation window.

Second, Patton and Sheppard (2015) employed the realized semivariances introduced by Barndorff-Nielsen, Kinnebrock, and Shephard (2010) in a semivariance HAR (SHAR) model by substituting the current-day realized variance by its up- and down-semivariance,

$$\log \widetilde{RV}_{i,t+1} = \beta_0 + \beta_d^+ \log \widetilde{RV}^+_{i,t} + \beta_d^- \log \widetilde{RV}^-_{i,t} + \beta_w \log \widetilde{RV}_{i,t-4:t} + \beta_m \log \widetilde{RV}_{i,t-21:t} + \widetilde{\xi}_{i,t}$$

with $\mathbf{E}[\widetilde{\xi}_{i,t}|\mathcal{F}_{t-1}] = 0$. Here, we once again use the log-values in contrast to Patton and Sheppard (2015) in order to get predictive values for the log-mean $\hat{\mu}^{SHAR}_{i,t+1|t}$ and log-standard deviation $\hat{\sigma}^{SHAR}_{i,t+1|t}$. We denote the one-step-ahead density forecast at day t with

$$f^{HAR}_{i,t+1|t} = f_{LNorm(\hat{\mu}^{HAR}_{i,t+1|t}, (\hat{\sigma}^{HAR}_{i,t+1|t})^2)} \text{ and } f^{SHAR}_{i,t+1|t} = f_{LNorm(\hat{\mu}^{SHAR}_{i,t+1|t}, (\hat{\sigma}^{SHAR}_{i,t+1|t})^2)},$$

where $f_{LNorm(\mu,\sigma^2)}$ is the log-normal density function with log-mean μ and log-variance σ^2. Both models are estimated using ordinary least squares estimation.

Both models are fitted separately for each stock on a daily rolling estimation window. The first estimation window starts in January 2000 and ends on the last day of December 2004. In total, we have $T = 4518$ days. The estimation window is of length $R = 1246$ and the out-of-sample period of length $P = 3272$.

3.4.3 Forecast evaluation for different volatility proxies

In order to examine the influence of using different volatility proxies for evaluating the distribution forecasts, we calculate the out-of-sample forecast errors both with respect to RV based on 5-minute returns and $\widetilde{\mathrm{RV}}$ based on 15-minute returns. A closed-form solution for the CRPS with respect to a log-normal distribution is derived in Baran and Lerch (2015). For the analytical solution for the QS we calculated the corresponding squared L^2-norm of a log-normal distribution $f_{LNorm(\mu,\sigma^2)}$ with log-mean μ and log-variance σ^2,

$$\|f_{LNorm(\mu,\sigma^2)}\|_2^2 = \frac{1}{2\sqrt{\pi}\sigma}\exp\left(\frac{\sigma^2}{4} - \mu\right).$$

A detailed derivation can be found in Appendix 3.6.2.

The forecast evaluation results are presented in Table 3.3. As in the simulation study, we consider the sensitivity of the DM test statistics for stock i using the two volatility proxies, denoted by t_i^{RV} and $t_i^{\widetilde{\mathrm{RV}}}$. In both cases, the variances of the corresponding loss differences are estimated using Newey-West standard errors.

The average DM test statistic for the log score and the 5-minute RV is 3.16 in comparison to an average value of 1.61 in the case of the CRPS as reported in column one of Table 3.3. We observe that the average difference in the test statistics is the lowest for the CRPS, mimicking our results in our previous simulation study. It is also noteworthy that the “bias” in the test statistic for the CRPS is almost only one-third of the value for the log score. As an additional measure to sensitivity of measurement errors, Table 3.3 also reports the standard deviation between the DM test statistic with respect to RV_t and $\widetilde{RV}_t$ in column three. The standard deviation of the difference across stocks is the highest for the log score and the lowest for the CRPS. Hence, we have further evidence that comparing distribution forecasts with respect to noisy proxies is less sensitive when using the CRPS instead of the log score.

3.5 Conclusion

This paper examines the evaluation of distribution forecasts in the presence of measurement error. First, we address the forecast ranking invariance under linear rescaling of the data; for example, reporting returns in percentages or annualized quarterly logarithmic growth rates. All commonly used loss functions are shown to imply the same expected ranking for the rescaled

Table 3.3: Out-of-sample forecast comparison for different volatility proxies.

	DM test stat.		
	$\overline{t_i^{\mathrm{RV}}}$	$\overline{t_i^{\widetilde{\mathrm{RV}}}} - \overline{t_i^{\mathrm{RV}}}$	$\mathrm{sd}({t_i}^{\widetilde{\mathrm{RV}}} - t_i^{\mathrm{RV}})$
LogS	3.16	3.75	1.71
QS	2.30	3.30	1.78
CRPS	1.61	**1.36**	**0.82**

Notes: In the first column we report the average t-statistic of the DM tests across 28 DJIA stocks when employing the more accurate volatility proxy based on 5-minute returns. The second and third column report the differences and standard deviations of the DM tests based on 5-minute and 15-minute returns. Numbers in bold report the smallest average difference and standard deviation. The initial estimation period comprises the data from 2000 to 2004 which also determines the length of the rolling estimation window. The losses are calculated on the evaluation period starting in 2005 and ending in 2017. DM tests are calculated using Newey-West standard errors.

data as they do for the original data. Second, we address the influence of additive measurement error present in many economic time series; for example, GDP growth and volatility. Evaluating distribution forecasts in the presence of these errors is particularly difficult. On the one hand, the forecasts are supposed to indicate the true uncertainty of the predictand's future outcomes but, on the other hand, they are evaluated against observations that are uncertain themselves.

Proper scoring rules will always favor forecasts that match the observations' distribution and not necessarily the distribution of the true predictand. However, in our theoretical findings we show that the quadratic score and the CRPS are less prone to change forecast rankings in the presence of observational error than the log score. Both the empirical application on a cross-section of 28 DJIA constituents and the simulation aligned with US GDP growth rates are in line with our theoretical results. The CRPS turns out to be the best measure for examining forecast performance in the presence of small and gross observational error. Even though the quadratic score is gross-error insensitive, the simulation study and the empirical findings suggests that it is only insensitive with respect to possibly large but less frequent observational errors. The CRPS does not suffer this drawback. Thus, it is our recommended forecast evaluation criterion.

As a direction for further research, we see the possible implications for multivariate forecasting. Evaluating multivariate predictive distributions makes it possible to assess both the forecast accuracy in each dimension and the joint dependency structure. It would be interesting to examine the sensitivity of joint density evaluation for a set of variables with a varying degree of measurement error. Empirical examples are the different degrees of measurement error in GDP and inflation or multivariate volatility forecasting for different stocks.

3.6 Appendix

3.6.1 Proofs

Lemma 3.1. *Let Y be a random variable with density function $f_Y(x)$, $f_Y \in L^\gamma$. Then, for the corresponding scaled random variable λY, $\lambda > 0$ it holds that $\|f_{\lambda Y}\|_\gamma = \lambda^{\frac{1-\gamma}{\gamma}} \|f_Y\|_\gamma$.*

Proof of Lemma 3.1. If $Y \sim f_Y$, then $f_{\lambda Y}(x) = f_Y(x/\lambda) \cdot 1/\lambda$. Hence, by substitution:

$$\begin{aligned}
\|f_{\lambda Y}\|_\gamma &= \left(\int_{-\infty}^{\infty} f_Y\left(\frac{x}{\lambda}\right)^\gamma \lambda^{-\gamma}\,\mathrm{d}x\right)^{1/\gamma} \\
&= \left(\lambda^{1-\gamma}\int_{-\infty}^{\infty} f_Y\left(\frac{x}{\lambda}\right)^\gamma \frac{1}{\lambda}\,\mathrm{d}x\right)^{1/\gamma} \\
&= \lambda^{\frac{1-\gamma}{\gamma}}\left(\int_{-\infty}^{\infty} f_Y(x)^\gamma\,\mathrm{d}x\right)^{1/\gamma} \\
&= \lambda^{\frac{1-\gamma}{\gamma}}\|f_Y\|_\gamma
\end{aligned}$$

□

Proof of Proposition 3.1. Let $Y \sim F_Y$ and consider the corresponding scaled variable λY, $\lambda > 0$. If Y has a continuous density f_Y, we have that $\lambda Y \sim f_{\lambda Y}(x) = f_Y(x/\lambda) \cdot 1/\lambda$.

LogS: The scaled log score is given by

$$\begin{aligned}
\mathrm{LogS}(f_{\lambda Y}, \lambda y) &= -\log(f_{\lambda Y}(\lambda y)) \\
&= -\log(f_Y(y)) + \log(\lambda) \\
&= \mathrm{LogS}(f_Y, y) + \log(\lambda)
\end{aligned}$$

Hence, it is scaling-invariant of order 0.

CLS: The initially defined region of interest $A \subseteq \mathbb{R}$ is scaled accordingly to $A_\lambda = \{x \in \mathbb{R} \mid x/\lambda \in A\}$. Applying our results for the log score, we have that

$$\begin{aligned}
\mathrm{CLS}(f_{\lambda Y}, \lambda y) &= -\mathbb{1}_{\{\lambda y \in A_\lambda\}} \log(f_{\lambda Y}(\lambda y)) - \mathbb{1}_{\{\lambda y \in A_\lambda^C\}} \log\left(\int_{A_\lambda^C} f_{\lambda Y}(s)\,\mathrm{d}s\right) \\
&= -\mathbb{1}_{\{\lambda y \in A_\lambda\}}[\log(f_Y(y)) - \log(\lambda)] - \mathbb{1}_{\{\lambda y \in A_\lambda^C\}} \log\left(\int_{A_\lambda^C} f_Y(s/\lambda)\cdot 1/\lambda\,\mathrm{d}s\right) \\
&= -\mathbb{1}_{\{y \in A\}} \log(f_Y(y)) + \mathbb{1}_{\{y \in A\}} \log(\lambda) - \mathbb{1}_{\{y \in A^C\}} \log\left(\int_{A^C} f_Y(s)\,\mathrm{d}s\right)
\end{aligned}$$

$$= \mathrm{CLS}(f_Y, y) + \mathbb{1}_{\{y \in A\}} \log(\lambda)$$

Hence, it is scaling-invariant of order 0.

PS: Next, we turn to the power score using Lemma 3.1:

$$\begin{aligned}
\mathrm{PS}_\gamma(f_{\lambda Y}, \lambda y) &= -\gamma f_{\lambda Y}(\lambda y)^{\gamma-1} + (\gamma-1)\|f_{\lambda Y}\|_\gamma^\gamma \\
&= -\gamma f_Y(y)^{\gamma-1}\lambda^{1-\gamma} + (\gamma-1)\lambda^{1-\gamma}\|f_Y\|_\gamma^\gamma \\
&= \lambda^{1-\gamma}\,\mathrm{PS}_\gamma(f_Y, y)
\end{aligned}$$

Therefore, the power score is scaling-invariant of order $1-\gamma$ and, thus, the specific case of the quadratic score is scaling-invariant of order -1.

PseudoS: Applying Lemma 3.1, we conclude for the pseudo spherical score the scaling-invariance of order $(1-\delta)/\delta$:

$$\begin{aligned}
\mathrm{PseudoS}_\delta(f_{\lambda Y}, \lambda y) &= -\frac{f_{\lambda Y}(\lambda y)^{\delta-1}}{\|f_{\lambda Y}\|_\delta^{\delta-1}} \\
&= -\lambda^{1-\delta} \cdot \lambda^{\frac{(1-\delta)^2}{\delta}} \cdot \frac{f_Y(y)^{\delta-1}}{\|f_Y\|_\delta^{\delta-1}} \\
&= \lambda^{\frac{1-\delta}{\delta}}\,\mathrm{PseudoS}_\delta(f_Y, y)
\end{aligned}$$

HyvS: The Hyvärinen score can be found to be scaling-invariant of order 2:

$$\begin{aligned}
\mathrm{HyvS}(f_{\lambda Y}, \lambda y) &= 2\frac{f''_{\lambda Y}(\lambda y)}{f_{\lambda Y}(\lambda y)} - \left(\frac{f'_{\lambda Y}(\lambda y)}{f_{\lambda Y}(\lambda x)}\right)^2 \\
&= 2\frac{f''_Y(y)\frac{1}{\lambda^3}}{f_Y(y)\frac{1}{\lambda}} - \left(\frac{f'_Y(x)\frac{1}{\lambda^2}}{f_Y(y)\frac{1}{\lambda}}\right)^2 \\
&= \frac{1}{\lambda^{-2}}\,\mathrm{HyvS}(f_Y, y)
\end{aligned}$$

CRPS: The CRPS is scaling invariant of order 1 by the change of variables formula:

$$\begin{aligned}
\mathrm{CRPS}(F_{\lambda Y}, \lambda y) &= \int_{-\infty}^{\infty} \left(F_{\lambda Y}(z) - \mathbb{1}_{\{z > \lambda y\}}(z)\right)^2 \,\mathrm{d}z \\
&= \int_{-\infty}^{\infty} \left(F_Y\left(\frac{z}{\lambda}\right) - \mathbb{1}_{\{z/\lambda > y\}}(z)\right)^2 \,\mathrm{d}z \\
&= \lambda \int_{-\infty}^{\infty} \left(F_Y(z) - \mathbb{1}_{\{z > y\}}(z)\right)^2 \,\mathrm{d}z
\end{aligned}$$

$$= \lambda \, \mathrm{CRPS}(F_Y, y)$$

□

Lemma 3.2. *Consider $Y \sim \mathcal{N}(0,1)$ and the Gaussian forecast density $f = \varphi_{\mu,\sigma^2}$, which is misspecified for $\mu \neq 0$ or $\sigma^2 \neq 1$. The corresponding expected log and continuous ranked probability score are given by*

$$\mathbf{E}_Y[\mathrm{LogS}(f,Y)] = \frac{1}{2}\log(2\pi) + \log\sigma + \frac{1}{2\sigma^2}(1+\mu^2)$$

and

$$\mathbf{E}_Y[\mathrm{CRPS}(f,Y)] = \sqrt{\frac{2(1+\sigma^2)}{\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2(1+\sigma^2)}\right) - 2\mu\Phi\left(\frac{-\mu}{\sqrt{1+\sigma^2}}\right) + \mu - \frac{\sigma}{\sqrt{\pi}}.$$

Proof. The expected LogS is given by

$$\begin{aligned}
\mathbf{E}_Y[\mathrm{LogS}(f,Y)] &= \mathbf{E}_Y\left[\frac{1}{2}\log(2\pi) + \log\sigma + \frac{1}{2}\left(\frac{Y-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{2}\log(2\pi) + \log\sigma + \frac{1}{2\sigma^2}\mathbf{E}_Y[Y^2 - 2Y\mu + \mu^2] \\
&= \frac{1}{2}\log(2\pi) + \log\sigma + \frac{1}{2\sigma^2}(1+\mu^2).
\end{aligned}$$

The continuous ranked probability score of the Gaussian distribution forecast f is given by (Gneiting et al., 2005):

$$\mathrm{CRPS}(f,y) = \sigma\left(\frac{y-\mu}{\sigma}\left(2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right) + 2\varphi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right)$$

Thus, for the expected continuous ranked probability score of a Gaussian distribution forecast we obtain with Stein's lemma:

$$\begin{aligned}
\mathbf{E}_Y[\mathrm{CRPS}(f,Y)] &= \sigma\mathbf{E}_Y\left[\frac{Y-\mu}{\sigma}\left(2\Phi\left(\frac{Y-\mu}{\sigma}\right) - 1\right) + 2\varphi\left(\frac{Y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right] \\
&= \mathbf{E}_Y\left[2Y\Phi\left(\frac{Y-\mu}{\sigma}\right) - 2\mu\Phi\left(\frac{Y-\mu}{\sigma}\right) - (Y-\mu) + 2\sigma\varphi\left(\frac{Y-\mu}{\sigma}\right) - \frac{\sigma}{\sqrt{\pi}}\right]
\end{aligned}$$

$$= \underbrace{2\,\mathbf{E}_Y\left[Y\Phi\left(\frac{Y-\mu}{\sigma}\right)\right]}_{=\mathbf{E}_Y[\varphi(\frac{Y-\mu}{\sigma})\frac{1}{\sigma}]} -2\mu\mathbf{E}_Y\left[\Phi\left(\frac{Y-\mu}{\sigma}\right)\right] + \mu + 2\sigma\mathbf{E}_Y\left[\varphi\left(\frac{Y-\mu}{\sigma}\right)\right] - \frac{\sigma}{\sqrt{\pi}}$$

$$= \frac{2(1+\sigma^2)}{\sigma}\mathbf{E}_Y\left[\varphi\left(\frac{Y-\mu}{\sigma}\right)\right] - 2\mu\mathbf{E}_Y\left[\Phi\left(\frac{Y-\mu}{\sigma}\right)\right] + \mu - \frac{\sigma}{\sqrt{\pi}}$$

Next, we calculate the two remaining expectations. First,

$$\begin{aligned}
\mathbf{E}_Y\left[\varphi\left(\frac{Y-\mu}{\sigma}\right)\right] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\,\mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}\left((1+\sigma^2)y^2 - 2y\mu + \mu^2\right)\right)\,\mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}\left((\sqrt{1+\sigma^2}y)^2 - \frac{2\sqrt{1+\sigma^2}y\mu}{\sqrt{1+\sigma^2}} + \frac{\mu^2}{1+\sigma^2} + \frac{\sigma^2\mu^2}{1+\sigma^2}\right)\right)\,\mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi(1+\sigma^2)}}\exp\left(-\frac{\mu^2}{2(1+\sigma^2)}\right) \\
&\quad\times\underbrace{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(\sqrt{1+\sigma^2}y - \frac{\mu}{\sqrt{1+\sigma^2}}\right)^2\right)\sqrt{1+\sigma^2}\,\mathrm{d}y}_{=1} \\
&= \frac{1}{\sqrt{2\pi(1+\sigma^2)}}\exp\left(-\frac{\mu^2}{2(1+\sigma^2)}\right)
\end{aligned}$$

Second, let $Z \sim \mathcal{N}(\mu,\sigma^2)$ independent of Y. Then, noting that $Z - Y \sim \mathcal{N}(\mu, 1+\sigma^2)$ and applying the law of total probability,

$$\begin{aligned}
\mathbf{E}_Y\left[\Phi\left(\frac{Y-\mu}{\sigma}\right)\right] &= \int_{-\infty}^{\infty}\Phi\left(\frac{y-\mu}{\sigma}\right)\,\mathrm{d}\,\mathbf{P}^Y(y) \\
&= \int_{-\infty}^{\infty}\mathbf{P}(Z\leq y)\,\mathrm{d}\,\mathbf{P}^Y(y) \\
&= \int_{-\infty}^{\infty}\mathbf{P}(Z\leq Y|Y=y)\,\mathrm{d}\,\mathbf{P}^Y(y) \\
&= \int_{-\infty}^{\infty}\mathbf{P}(Z-Y\leq 0|Y=y)\,\mathrm{d}\,\mathbf{P}^Y(y) \\
&= \mathbf{P}(Z-Y\leq 0) \\
&= \Phi\left(\frac{-\mu}{\sqrt{1+\sigma^2}}\right)
\end{aligned}$$

Jointly,

$$\mathbf{E}_Y[\mathrm{CRPS}(f,Y)] = \sqrt{\frac{2(1+\sigma^2)}{\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2(1+\sigma^2)}\right) - 2\mu\Phi\left(\frac{-\mu}{\sqrt{1+\sigma^2}}\right) + \mu - \frac{\sigma}{\sqrt{\pi}}.$$

□

Proof of Proposition 3.2. (a) If L is scaling-invariant, we have that for all $\lambda > 0$

$$\begin{aligned}\mathbf{E}_Y[L(F_\lambda,\lambda Y) - L(G_\lambda,\lambda Y)] \geq 0 &\Leftrightarrow \lambda^k \mathbf{E}[L(F,Y) - L(G,Y)] \geq 0 \\ &\Leftrightarrow \mathbf{E}[L(F,Y) - L(G,Y)] \geq 0\end{aligned}$$

(b) Define $\widetilde{S}(f,y) = \mathrm{LogS}(f,y) + \mathrm{CRPS}(f,y)$. By construction, the loss function $\widetilde{S}$ is not scaling-invariant by being a sum of scaling-invariant loss functions of different order. Now, let $Y \sim \mathcal{N}(0,1)$. We calculate the expected score difference of two misspecified Gaussian forecasts $g(x) = \varphi_{0,2}(x)$ and $\widetilde{g}(x) = \varphi_{2,1/2}(x)$ using Lemma 3.2 and obtain:

$$\begin{aligned}\mathbf{E}_Y[\widetilde{S}(g,Y) - \widetilde{S}(\widetilde{g},Y)] &= \mathbf{E}_Y[\mathrm{LogS}(g,Y) - \mathrm{LogS}(\widetilde{g},Y)] + \mathbf{E}_Y[\mathrm{CRPS}(g,Y) - \mathrm{CRPS}(\widetilde{g},Y)] \\ &\approx -8.49 + 1.54 < 0\end{aligned}$$

Thus, g is preferred over $\widetilde{g}$ with respect to $\widetilde{S}$. On the other hand, looking at scaled observations λY, $\lambda > 0$ and the corresponding scaled density forecasts g_λ and $\widetilde{g}_\lambda$, Proposition 3.1 implies

$$\begin{aligned}&\mathbf{E}_Y[\widetilde{S}(f_\lambda,\lambda Y) - \widetilde{S}(g_\lambda,\lambda Y)] \\ &\quad= \mathbf{E}_Y[\mathrm{LogS}(g_\lambda,\lambda Y) - \mathrm{LogS}(\widetilde{g}_\lambda,\lambda Y)] + \mathbf{E}_Y[\mathrm{CRPS}(g_\lambda,\lambda Y) - \mathrm{CRPS}(\widetilde{g}_\lambda,\lambda Y)] \\ &\quad\approx -8.49 + \lambda \cdot 1.54,\end{aligned}$$

which may be greater than zero, for example if $\lambda = 100$; that is, presenting numbers in percentages. In this case, rescaling the data changes the order of expected losses.

□

Proof of Proposition 3.3. Let Ω_G denote the support of the error distribution.

(a) Let f be a standard Gaussian density and $U \sim G$ such that $\Omega_G = \mathbb{R}$ and $\mathbf{E}_U[U^2] < \infty$.

Then, applying the reverse triangle inequality,

$$\begin{aligned}
\mathrm{EADF}(\mathrm{LogS}, f, y, G) &= \mathbf{E}_U \left| \frac{(y+U)^2}{2} - \frac{y^2}{2} \right| \\
&= \frac{1}{2}\mathbf{E}_U \left| 2yU + U^2 \right| \\
&= \frac{1}{2} \int_{\mathbb{R}} |2yu - (-u^2)| \; \mathrm{d}G(u) \\
&\geq \frac{1}{2} \int_{\mathbb{R}} \Big| |2yu| - |u^2| \Big| \; \mathrm{d}G(u) \\
&= \frac{1}{2} \int_{\mathbb{R}} |u| \cdot \big| |2y| - |u| \big| \; \mathrm{d}G(u) \\
&\geq \frac{1}{2} \int_{u \in \mathbb{R}:\ |2y|>|u|} |u| \cdot \big(|2y| - |u| \big) \; \mathrm{d}G(u) \\
&= |y| \underbrace{\int_{u \in \mathbb{R}:\ |2y|>|u|} |u| \; \mathrm{d}G(u)}_{\to \mathbf{E}_U|U| \ as\ |y| \to \infty} - \frac{1}{2} \underbrace{\int_{u \in \mathbb{R}:\ |2y|>|u|} u^2 \; \mathrm{d}G(u)}_{\to \mathbf{E}_U[U^2] \ as\ |y| \to \infty} .
\end{aligned}$$

Hence, $\gamma^* = \infty$ as the last expression increases in $|y|$ without boundary.

(b) If $f(x) \leq \overline{f}$, we have that $|f(x_1) - f(x_2)| \leq \overline{f}$ for all x_1, x_2 in the support of f. Hence,

$$\begin{aligned}
\mathrm{EADF}(\mathrm{QS}, f, y, G) &= \mathbf{E}_U \left| -2f(y+U) + \|f\|_2^2 - \Big(-2f(y) + \|f\|_2^2 \Big) \right| \\
&= 2\mathbf{E}_U \left| f(y+U) - f(y) \right| \\
&\leq 2\overline{f},
\end{aligned}$$

which implies $\gamma^* < \infty$.

(c) Let $Y, Y' \sim F$ be independent and $U \sim G$ with $\mathbf{E}|U| < \infty$. Using the kernel score representation of the continuous ranked probability score (Gneiting and Raftery, 2007),

$$\mathrm{CRPS}(F, y) = \mathbf{E}_Y |Y - y| - \frac{1}{2}\mathbf{E}_{Y,Y'} |Y - Y'|,$$

which is valid for any distribution forecast F with finite first moment, we derive

$$\begin{aligned}
\mathrm{EADF}(\mathrm{CRPS}, F, y, G) &= \mathbf{E}_U \left| \mathrm{CRPS}(F, y+U) - \mathrm{CRPS}(F, y) \right| \\
&= \mathbf{E}_U \left| \mathbf{E}_Y |Y - y - U| - \mathbf{E}_Y |Y - y| \right| \\
&= \int_{\Omega_G} \left| \int_{\Omega_F} |x - y - u| dF(x) - \int_{\Omega_F} |x - y| \; \mathrm{d}F(x) \right| \; \mathrm{d}G(u)
\end{aligned}$$

$$\begin{aligned}
&\leq \int_{\Omega_G} \int_{\Omega_F} \big| |x-y-u| - |x-y| \big| \;\mathrm{d}F(x)\;\mathrm{d}G(u) \\
&\leq \int_{\Omega_G} \int_{\Omega_F} |u| \;\mathrm{d}F(x)\;\mathrm{d}G(u) \\
&= \mathbf{E}_U \, |U| \,.
\end{aligned}$$

as $||a-b|-|a|| \leq |b|$ for all $a, b \in \mathbb{R}$. This implies $\gamma^* < \infty$.

□

3.6.2 Analytical expressions of the quadratic and power score

Closed-form solutions of the quadratic score for the normal and the log-normal distribution are used in this paper. Beyond these, we provide additional results for the mixture of normals, student-t, generalized beta, and two-piece normal distribution.

Note that the analytical solutions for the L^γ norm can also be used to calculate analytical expressions for the pseudo spherical score.

Normal

Proposition 3.4. *Let $\gamma > 1$. The power score of degree γ with respect to a normal distribution density forecast $f_{Norm(\mu,\sigma^2)}$ with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$,*

$$f_{Norm(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right),$$

is given by

$$\mathrm{PS}_\gamma(f_{Norm(\mu,\sigma^2)}, y) = -\frac{\gamma}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\gamma-1)(y-\mu)^2}{2\sigma^2}\right) + \frac{\gamma-1}{\sqrt{\gamma}(2\pi)^{\frac{\gamma-1}{2}}\sigma^{\gamma-1}},$$

which implies

$$\mathrm{QS}(f_{Norm(\mu,\sigma^2)}, y) = -\frac{\sqrt{2}}{\sqrt{\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) + \frac{1}{2\sqrt{\pi}\sigma}.$$

Proof. Straightforward calculations yield

$$\begin{aligned}
\|f_{Norm(\mu,\,\sigma^2)}\|_\gamma^\gamma &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)\right)^\gamma \;\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\gamma/2}\sigma^\gamma} \exp\left(-\frac{\gamma}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \;\mathrm{d}x
\end{aligned}$$

$$= \frac{1}{(2\pi)^{\gamma/2}\sigma^\gamma} \frac{\sqrt{2\pi}\sigma}{\sqrt{\gamma}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{\gamma}}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\left(\frac{\sigma}{\sqrt{\gamma}}\right)^2}\right) \mathrm{d}x}_{=1}$$

$$= \frac{1}{\sqrt{\gamma}(2\pi)^{\frac{\gamma-1}{2}}\sigma^{\gamma-1}}.$$

□

Mixture of normals

For the case of the mixture of normals, we only report an analytical solution for the quadratic score and not for a general power score of order $\gamma > 1$. A generalization to arbitrary $\gamma > 1$ may be possible but would possibly involve Newton's generalized binomial theorem and integrals of infinite series. We leave this for future research.

Proposition 3.5. *Let $f_{MixNorm(\mu,\sigma^2)}$, $\mu \in \mathbb{R}^n, \sigma \in \mathbb{R}_+^n$, be a density of a mixture of $n \in \mathbb{N}$ normal distributions, $f(x) = \sum_{i=1}^n w_i\varphi_i(x)$ with weights $w_i \geq 0$, $\sum_{i=1}^n w_i = 1$, and where $\varphi_i(x)$ denote normal densities with individual mean μ_i and standard deviation $\sigma_i > 0$. Then, the quadratic score is given by*

$$\begin{aligned}\mathrm{QS}(f_{MixNorm(\mu,\sigma^2)}, y) = -2f_{MixNorm(\mu,\sigma^2)} + \sum_{i=1}^n \frac{w_i^2}{2\sqrt{\pi}\sigma_i} \\ + \sum_{\substack{i,j=1 \\ i\neq j}}^n w_iw_j \frac{1}{\sqrt{2\pi(\sigma_i^2+\sigma_j^2)}} \exp\left(-\frac{(\mu_i-\mu_j)^2}{2(\sigma_i^2+\sigma_j^2)}\right).\end{aligned}$$

Proof. We calculate the squared L^2 norm of the respective density:

$$\begin{aligned}\|f_{MixNorm(\mu,\sigma^2)}\|_2^2 &= \|\sum_{i=1}^n w_i\varphi_i\|_2^2 \\ &= \int_{-\infty}^{\infty}\left(\sum_{i=1}^n w_i\varphi_i(x)\right)^2 \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \sum_{i=1}^n w_i^2\varphi_i(x)^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^n w_iw_j\varphi_i(x)\varphi_j(x)\ \mathrm{d}x \\ &= \sum_{i=1}^n w_i^2\|\varphi_i\|_2^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^n w_iw_j \int_{-\infty}^{\infty} \varphi_i(x)\varphi_j(x)\ \mathrm{d}x\end{aligned}$$

$$= \sum_{i=1}^{n} \frac{w_i^2}{2\sqrt{\pi}\sigma_i} + \sum_{\substack{i,j=1\\ i\neq j}}^{n} w_i w_j \frac{1}{\sqrt{2\pi(\sigma_i^2+\sigma_j^2)}} \exp\left(-\frac{(\mu_i-\mu_j)^2}{2(\sigma_i^2+\sigma_j^2)}\right),$$

where the last equation follow from Proposition 3.4 and completing the square,

$$\int_{-\infty}^{\infty} \varphi_i(x)\varphi_j(x)\ \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_i\sigma_j} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu_i)^2}{\sigma_i^2} + \frac{(x-\mu_j)^2}{\sigma_j^2}\right)\right)\ \mathrm{d}x$$

$$= \frac{1}{\sqrt{2\pi(\sigma_i^2+\sigma_j^2)}} \int_{-\infty}^{\infty} \frac{\sqrt{\sigma_i^2+\sigma_j^2}}{\sqrt{2\pi}\sigma_i\sigma_j} \exp\left(-\frac{1}{2}\left(\frac{\left(x-\frac{\mu_i\sigma_j^2+\mu_j\sigma_i^2}{\sigma_i^2+\sigma_j^2}\right)^2 + \frac{\sigma_i\sigma_j(\mu_i-\mu_j)^2}{(\sigma_i^2+\sigma_j^2)^2}}{\sigma_i^2\sigma_j^2(\sigma_i^2+\sigma_j^2)^{-1}}\right)\right)\ \mathrm{d}x$$

$$= \frac{1}{\sqrt{2\pi(\sigma_i^2+\sigma_j^2)}} \exp\left(-\frac{(\mu_i-\mu_j)^2}{2(\sigma_i^2+\sigma_j^2)}\right).$$

□

Log-normal

Proposition 3.6. *Let $\gamma > 1$. The power score of degree γ with respect to a log-normal distribution density forecast $f_{LNorm(\mu,\sigma^2)}$ with parameters μ and $\sigma > 0$,*

$$f_{LNorm(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\frac{(\log x-\mu)^2}{\sigma^2}\right), x > 0,$$

is given by

$$\mathrm{PS}_\gamma(f_{LNorm(\mu,\sigma^2)}, y) = -\gamma f_{LNorm(\mu,\sigma^2)}(y)^{\gamma-1} + \frac{(\gamma-1)\exp\left(\frac{1}{2\gamma}((\gamma-1)^2\sigma^2 - 2\gamma\mu(\gamma-1))\right)}{(2\pi)^{\gamma/2-1}\sigma^{\gamma-1}\sqrt{\gamma}}$$

which implies

$$\mathrm{QS}(f_{LNorm(\mu,\sigma^2)}, y) = -2f_{LNorm(\mu,\sigma^2)}(x) + \frac{1}{2\sqrt{\pi}\sigma} \exp\left(\frac{\sigma^2}{4} - \mu\right).$$

Proof. The L^γ norm can be calculated as

$$\|f_{LNorm(\mu,\,\sigma^2)}\|_\gamma^\gamma = \int_0^\infty \left(\frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\frac{(\log x-\mu)^2}{\sigma^2}\right)\right)^\gamma\ \mathrm{d}x$$

$$
\begin{aligned}
&= \int_0^\infty \frac{1}{(2\pi)^{\gamma/2}\sigma^\gamma x^\gamma} \exp\left(-\frac{\gamma}{2}\frac{(\log x - \mu)^2}{\sigma^2}\right) \,\mathrm{d}x \\
&= \int_{-\infty}^\infty \frac{1}{(2\pi)^{\gamma/2}\sigma^\gamma e^{(\gamma-1)u}} \exp\left(-\frac{\gamma}{2}\frac{(u - \mu)^2}{\sigma^2}\right) \,\mathrm{d}u \\
&= \int_{-\infty}^\infty \frac{1}{(2\pi)^{\gamma/2}\sigma^\gamma} \exp\left(-\frac{\gamma}{2}\frac{(u - \mu)^2 + 2\sigma^2(\gamma-1)u/\gamma}{\sigma^2}\right) \,\mathrm{d}u \\
&= \int_{-\infty}^\infty \frac{1}{(2\pi)^{\gamma/2}\sigma^\gamma} \\
&\quad \times \exp\left(-\frac{\gamma}{2}\frac{u^2 - 2u(\gamma\mu - (\gamma-1)\sigma^2)/\gamma + (\gamma\mu - (\gamma-1)\sigma^2)^2/\gamma^2}{\sigma^2}\right) \\
&\quad \times \exp\left(-\frac{\gamma}{2}\frac{-(\gamma\mu - (\gamma-1)\sigma^2)^2/\gamma^2 + \mu^2}{\sigma^2}\right) \,\mathrm{d}u \\
&= \frac{\exp\left(\frac{1}{2\gamma}((\gamma-1)^2\sigma^2 - 2\gamma\mu(\gamma-1))\right)}{(2\pi)^{\gamma/2-1}\sigma^{\gamma-1}\sqrt{\gamma}} \\
&\quad \times \underbrace{\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma/\sqrt{\gamma}} \exp\left(-\frac{1}{2}\frac{(u - (\gamma\mu - (\gamma-1)\sigma^2)/\gamma)^2}{\sigma^2/\gamma}\right) \,\mathrm{d}u}_{=1} \\
&= \frac{\exp\left(\frac{1}{2\gamma}((\gamma-1)^2\sigma^2 - 2\gamma\mu(\gamma-1))\right)}{(2\pi)^{\gamma/2-1}\sigma^{\gamma-1}\sqrt{\gamma}}.
\end{aligned}
$$

In the case of the L^2 norm it simplifies to

$$\|f_{LNorm(\mu,\sigma^2)}\|_2^2 = \frac{1}{2\sqrt{\pi}\sigma}\exp\left(\frac{\sigma^2}{4} - \mu\right).$$

□

Student-t

For calculating $\|\cdot\|_\gamma^\gamma$ of a student-t distribution with $\nu > 0$ degrees of freedom, we prove the following lemma.

Lemma 3.3. *Let $\gamma > 1$ and $\nu > 0$. It holds that*

$$\int_0^\infty \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\gamma(\nu+1)}{2}} \mathrm{d}x = \frac{\sqrt{\nu\pi}\,\Gamma\left(\frac{\gamma(\nu+1)-1}{2}\right)}{2\Gamma\left(\frac{\gamma(\nu+1)}{2}\right)}. \tag{3.7}$$

where Γ denotes the Gamma function.

Proof. Calculating the left-hand integral by substitution $x = \sqrt{\nu}\tan(u)$, which is invertible over

$0 < u < \frac{\pi}{2}$ with inverse $u = \tan^{-1}(x/\sqrt{\nu})$),

$$\begin{aligned}\int_0^\infty \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{\frac{\gamma(\nu+1)}{2}}}\,\mathrm{d}x &= \int_0^{\frac{\pi}{2}} \frac{1}{\left(1+\tan^2(u)\right)^{\frac{\gamma(\nu+1)}{2}}}\sqrt{\nu}\sec^2(u)\,\mathrm{d}u \\ &= \sqrt{\nu}\int_0^{\frac{\pi}{2}} \frac{1}{\left(\sec^2(u)\right)^{\frac{\gamma(\nu+1)}{2}}}\sec^2(u)\,\mathrm{d}u \\ &= \sqrt{\nu}\int_0^{\frac{\pi}{2}} \cos^{\gamma(\nu+1)-2}(u)\,\mathrm{d}u,\end{aligned}$$

in which we use that $1+\tan^2(u) = \sec^2(u)$ and $\sec^{-1}(u) = \cos(u)$. Next, we calculate the remaining right-hand integral:

$$\begin{aligned}\int_0^{\frac{\pi}{2}} \cos^{\gamma(\nu+1)-2}(u)du &= \int_0^{\frac{\pi}{2}} \sin^{2\cdot\frac{1}{2}-1}(u)\cdot\cos^{2\left(\frac{\gamma(\nu+1)-1}{2}\right)-1}(u)\,\mathrm{d}u \\ &= \frac{1}{2}B\left(\frac{1}{2},\frac{\gamma(\nu+1)-1}{2}\right) \\ &= \frac{1}{2}\cdot\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\gamma(\nu+1)-1}{2}\right)}{\Gamma\left(\frac{\gamma(\nu+1)}{2}\right)} \\ &= \frac{\sqrt{\pi}\Gamma\left(\frac{\gamma(\nu+1)-1}{2}\right)}{2\Gamma\left(\frac{\gamma(\nu+1)}{2}\right)}\end{aligned}$$

as $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ which implies Equation (3.7). □

Proposition 3.7. *The power score of degree $\gamma > 1$ with respect to a student-t distribution with $\nu > 0$ degrees of freedom,*

$$f_{\textit{stud-t}(\nu)}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}}$$

is given by

$$\mathrm{PS}_\gamma(f_{\textit{stud-t}(\nu)}, y) = -\gamma f_{\textit{stud-t}(\nu)}(y)^{\gamma-1} + (\gamma-1)(\nu\pi)^{\frac{1-\gamma}{2}}\frac{\Gamma^\gamma\left(\frac{\nu+1}{2}\right)}{\Gamma^\gamma\left(\frac{\nu}{2}\right)}\frac{\Gamma\left(\frac{\gamma(\nu+1)-1}{2}\right)}{\Gamma\left(\frac{\gamma(\nu+1)}{2}\right)} \tag{3.8}$$

and, thus, the corresponding quadratic score for $\gamma = 2$ is given by

$$\mathrm{QS}(f_{\textit{stud-t}(\nu)}, y) = -2f_{\textit{stud-t}(\nu)}(y) + \frac{\Gamma^2\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\Gamma^2\left(\frac{\nu}{2}\right)}\frac{\Gamma\left(\nu+\frac{1}{2}\right)}{\sqrt{\pi}\Gamma\left(\nu+1\right)}. \tag{3.9}$$

Proof. Equation (3.7) in Lemma 3.3 readily implies

$$
\begin{aligned}
\|f_{\textit{stud-t}}\|_\gamma^\gamma &= \int_{-\infty}^{\infty} \left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}} \right)^\gamma \mathrm{d}x \\
&= \frac{\Gamma^\gamma\left(\frac{\nu+1}{2}\right)}{(\nu\pi)^{\gamma/2}\Gamma^\gamma\left(\frac{\nu}{2}\right)} \int_{-\infty}^{\infty} \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{\frac{\gamma(\nu+1)}{2}}} \mathrm{d}x \\
&= \frac{\Gamma^\gamma\left(\frac{\nu+1}{2}\right)}{(\nu\pi)^{\gamma/2}\Gamma^\gamma\left(\frac{\nu}{2}\right)} \cdot 2 \int_{0}^{\infty} \left(1+\frac{x^2}{\nu}\right)^{-\frac{\gamma(\nu+1)}{2}} \mathrm{d}x \\
&= \frac{\Gamma^\gamma\left(\frac{\nu+1}{2}\right)}{(\nu\pi)^{\gamma/2}\Gamma^\gamma\left(\frac{\nu}{2}\right)} \cdot \frac{\sqrt{\nu\pi}\Gamma\left(\frac{\gamma(\nu+1)-1}{2}\right)}{\Gamma\left(\frac{\gamma(\nu+1)}{2}\right)} \\
&= (\nu\pi)^{\frac{1-\gamma}{2}} \frac{\Gamma^\gamma\left(\frac{\nu+1}{2}\right)}{\Gamma^\gamma\left(\frac{\nu}{2}\right)} \frac{\Gamma\left(\frac{\gamma(\nu+1)-1}{2}\right)}{\Gamma\left(\frac{\gamma(\nu+1)}{2}\right)}.
\end{aligned}
$$

Hence, for $\gamma = 2$ we have that

$$
\|f_{\textit{stud-t}}\|_2^2 = \frac{\Gamma^2\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\Gamma^2\left(\frac{\nu}{2}\right)} \frac{\Gamma\left(\nu+\frac{1}{2}\right)}{\sqrt{\pi}\Gamma\left(\nu+1\right)}.
$$

□

Remark 3.3. *For ever increasing degrees of freedom, the student-t distribution approaches the standard normal distribution. The consistency of our results regarding the quadratic score in Proposition 3.7 and Proposition 3.4 for $\nu \to \infty$ follows from Wendel's limit (Wendel, 1948),*

$$
\lim_{x\to\infty} x^{-s} \frac{\Gamma(x+s)}{\Gamma(x)} = 1,
$$

for arbitrary real numbers s and x. This equality implies

$$
\lim_{x\to\infty} x^{t-s} \frac{\Gamma(x+s)}{\Gamma(x+t)} = \lim_{x\to\infty} x^{-s} \frac{\Gamma(x+s)}{\Gamma(x)} x^{t} \frac{\Gamma(x)}{\Gamma(x+t)} = 1
$$

for real numbers s, t and x, which leads us to

$$
\lim_{\nu\to\infty} \frac{\Gamma^2\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\Gamma^2\left(\frac{\nu}{2}\right)} \cdot \frac{\Gamma\left(\nu+\frac{1}{2}\right)}{\sqrt{\pi}\Gamma\left(\nu+1\right)}
$$

$$= \lim_{\nu\to\infty} \frac{1}{2} \cdot \underbrace{\frac{2\Gamma^2\left(\frac{\nu+1}{2}\right)}{\nu\Gamma^2\left(\frac{\nu}{2}\right)}}_{\to 1} \cdot \underbrace{\frac{\sqrt{\nu}\Gamma\left(\nu+\frac{1}{2}\right)}{\Gamma\left(\nu+1\right)}}_{\to 1} \cdot \frac{1}{\sqrt{\pi}}$$
$$= \frac{1}{2\sqrt{\pi}}.$$

Generalized beta distribution

Proposition 3.8. *The power score of degree $\gamma > 1$ with respect to a generalized beta density forecast*

$$f_{GBeta(a,b,l,r)}(x) = \frac{(x-l)^{a-1}(r-x)^{b-1}}{B(a,b)(r-l)^{a+b-1}},$$

shape parameter $a, b > \frac{1}{\gamma}$, and upper and lower bound $l, r \in \mathbb{R}, l < r$, is given by

$$\mathrm{PS}_\gamma(f_{GBeta(a,\,b,\,l,\,r)}, y) = \begin{cases} -\gamma f_{GBeta(a,b,l,r)}(y)^{\gamma-1} + \frac{(\gamma-1)B(\gamma(a-1)+1,\gamma(b-1)+1)}{(r-l)^{\gamma-1}B(a,b)^\gamma} & \textit{if } l \le y \le r, \\ \frac{(\gamma-1)B(\gamma(a-1)+1,\gamma(b-1)+1)}{(r-l)^{\gamma-1}B(a,b)^\gamma} & \textit{else,} \end{cases}$$

which implies

$$QS(f_{GBeta(a,\,b,\,l,\,r)}, y) = \begin{cases} -2f_{GBeta(a,b,l,r)}(y) + \frac{B(2a-1,2b-1)}{(r-l)B(a,b)^2} & \textit{if } l \le y \le r, \\ \frac{B(2a-1,2b-1)}{(r-l)B(a,b)^2} & \textit{else.} \end{cases}$$

Proof. By rearranging terms we conclude,

$$\begin{aligned} \|f_{GBeta(a,b,l,r)}\|_\gamma^\gamma &= \int_l^r \left(\frac{1}{B(a,b)(r-l)^{a+b-1}}(x-l)^{a-1}(r-x)^{b-1}\right)^\gamma \,\mathrm{d}x \\ &= \frac{1}{B(a,b)^\gamma(r-l)^{(a+b-1)\gamma}} \int_l^r (x-l)^{\gamma(a-1)}(r-x)^{\gamma(b-1)} \,\mathrm{d}x \\ &= \frac{B(\gamma(a-1)+1,\gamma(b-1)+1)\cdot(r-l)^{\gamma(a+b-1)-\gamma+1}}{B(a,b)^\gamma(r-l)^{(a+b-1)\gamma}} \\ &\quad \times \underbrace{\int_l^r \frac{(x-l)^{\gamma(a-1)+1-1}(r-x)^{\gamma(b-1)+1-1}}{B(\gamma(a-1)+1,\gamma(b-1)+1)\cdot(r-l)^{\gamma(a+b-1)-\gamma+1}} \,\mathrm{d}x}_{=1} \\ &= \frac{B(\gamma(a-1)+1,\gamma(b-1)+1)}{(r-l)^{\gamma-1}B(a,b)^\gamma}. \end{aligned}$$

□

Remark 3.4. *The constrained parameter space assumption $a, b > 1/\gamma$ in Proposition 3.8 is* *necessary for* $\|f_{GBeta(a,b,l,r)}\|_\gamma^\gamma$ *to be defined. From a practical point of view this is no restraint in most applications as it is often assumed that* $a, b > 1$ *in order to ensure unimodality and continuity at* $x = l$ *and* $x = r$*.*

Two-piece normal

Next, we consider the two-piece normal distribution of Wallis (2004, 2014).

Proposition 3.9. *Let* $\gamma > 1$*. The power score of degree* γ *with respect to a two-piece normal distribution density forecast* $f_{TPN(\mu_1,\mu_2,\sigma_1,\sigma_2)}$ *with location parameters* $\mu \in \mathbb{R}$ *and scale parameters* $\sigma_1, \sigma_2 > 0$*,*

$$f_{TPN(\mu,\sigma_1,\sigma_2)}(y) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_1^2}\right) & \text{if } y \leq \mu, \\ \frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_2^2}\right) & \text{if } y \geq \mu. \end{cases}$$

is given by

$$\mathrm{PS}_\gamma(f_{TPN(\mu,\sigma_1^2,\sigma_2^2)}, y) = -\gamma f^{\gamma-1}_{TPN(\mu,\sigma_1^2,\sigma_2^2)}(y) + \frac{(\gamma-1)2^{(\gamma-1)/2}}{\pi^{(\gamma-1)/2}(\sigma_1+\sigma_2)^{\gamma-1}\sqrt{\gamma}}$$

which implies

$$\mathrm{QS}(f_{TPN}, y) = -2f_{TPN}(y) + \frac{1}{\pi(\sigma_1+\sigma_2)}.$$

Proof. We calculate the L^γ norm to the power of γ,

$$\begin{aligned} \|f_{TPN(\mu,\sigma_1,\sigma_2)}\|_\gamma^\gamma &= \int_{-\infty}^{\infty} f_{TPN(\mu,\sigma_1,\sigma_2)}(x)^\gamma \,\mathrm{d}x \\ &= \int_{-\infty}^{\mu} \left(\frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_1^2}\right)\right)^\gamma \mathrm{d}x \\ &\quad + \int_{\mu}^{\infty} \left(\frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_2^2}\right)\right)^\gamma \mathrm{d}x. \end{aligned}$$

Next, we turn to the left-hand expression,

$$\begin{aligned} &\int_{-\infty}^{\mu} \left(\frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_1^2}\right)\right)^\gamma \mathrm{d}x \\ &= \frac{2^{\gamma/2}}{\pi^{\gamma/2}(\sigma_1+\sigma_2)^\gamma} \int_{-\infty}^{\mu} \exp\left(-\frac{\gamma}{2}\frac{(x-\mu)^2}{\sigma_1^2}\right) \mathrm{d}x \end{aligned}$$

$$= \frac{2^{\gamma/2}}{\pi^{\gamma/2}(\sigma_1+\sigma_2)^{\gamma}} \frac{\sqrt{2\pi}\sigma_1}{\sqrt{\gamma}} \underbrace{\int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi}\frac{\sigma_1}{\sqrt{\gamma}}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\left(\frac{\sigma_1}{\sqrt{\gamma}}\right)^2} \right) \mathrm{d}x}_{=\frac{1}{2}}$$

$$= \frac{2^{(\gamma-1)/2}\sigma_1}{\pi^{(\gamma-1)/2}(\sigma_1+\sigma_2)^{\gamma}\sqrt{\gamma}}.$$

Thus,

$$\begin{aligned} \|f_{TPN}\|_{\gamma}^{\gamma} &= \frac{2^{(\gamma-1)/2}\sigma_1}{\pi^{(\gamma-1)/2}(\sigma_1+\sigma_2)^{\gamma}\sqrt{\gamma}} + \frac{2^{(\gamma-1)/2}\sigma_2}{\pi^{(\gamma-1)/2}(\sigma_1+\sigma_2)^{\gamma}\sqrt{\gamma}} \\ &= \frac{2^{(\gamma-1)/2}}{\pi^{(\gamma-1)/2}(\sigma_1+\sigma_2)^{\gamma-1}\sqrt{\gamma}}. \end{aligned}$$

□

3.6.3 Data and revisions in US GDP growth.

For calibrating our simulation, we obtain quarterly releases of seasonally adjusted real GDP estimates (GDP_t^k) from the Federal Reserve Bank of Philadelphia.[8] The subscript t refers to the time period of the observation of GDP and the superscript k to the release wave.[9] The first release ($k = 1$) for quarter t is published in the following quarter $t + 1$. We obtained observations for the range from 1966Q1 to 2015Q4 and their corresponding first release until the release issued twelve quarters later. The observation for $k = 12$ will be considered to be the “final release” (e.g., Aruoba, 2008; Jacobs and Van Norden, 2011). The lag in releases of GDP estimates implies that the last observation, the final release for 2015Q4, is announced in 2019Q1.

Based on GDP_t^k, we calculate annualized quarter-over-quarter log GDP growth rates in percentages; that is, $\Delta GDP_t^k = 400 \times (\log GDP_t^k - \log GDP_{t-1}^{k+1})$. The increase from k to $k+1$ when calculating the quarterly growth rates is due to the fact that at a time where there is the k-th release of GDP_t, we already know the $(k+1)$-th release of GDP_{t-1}.

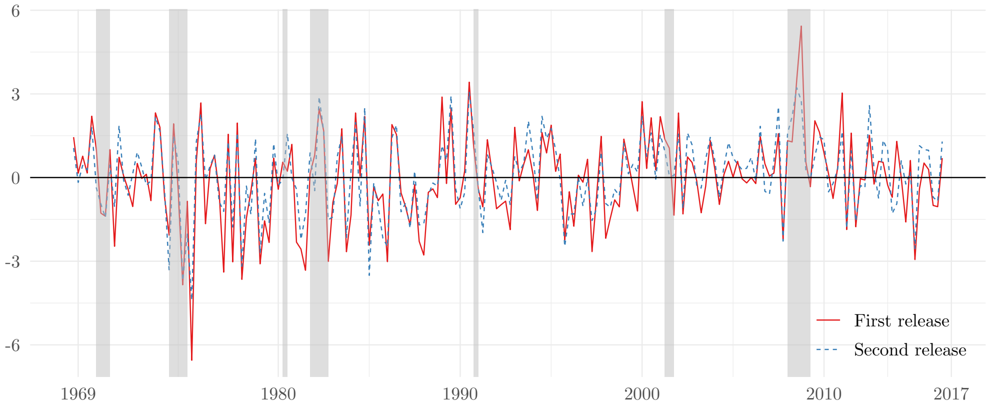
In Figure 3.3, we depict the first and second release errors over time. In general, there seems to be no clear pattern in the observation error. Two observations that stand out are the largest upward revision in GDP growth at the end of the first oil crisis in the 1970s and the largest

[8] https://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/data-files/routput

[9] Technically, our series of “GDP” is partially based on gross national product.

downward revision during the financial crisis in the 2000s. Additionally, it may be said that the revision error variance slightly decreased in the second half of the sample. This possible regime change amidst the great moderation is our reason to restrict ourselves to a sample starting in 1989Q1 for the calibration of our simulation. A natural question that arises is whether measurement errors in GDP are predictable or not. Among others, Aruoba (2008) document evidence of biasedness and predictability in multiple releases of macroeconomic variables. In contrast, Faust, Rogers, and Wright (2005) find no predictability in US GDP revisions which we will count as evidence for a mean-zero noise model. Similarly, Jacobs and Van Norden (2011) show that the US Bureau of Economic Analysis does incorporate all relevant information for each release.

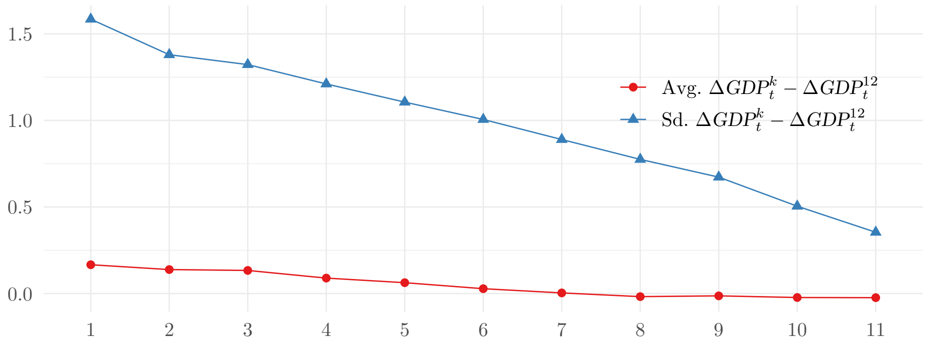
Figure 3.3: Observation error $\Delta GDP_t^k - \Delta GDP_t^{12}$ for the first $(k = 1)$ and second release $(k = 2)$ of GDP growth.



Notes: We depict quarterly growth rate errors between the first/second release of GDP data and the corresponding "final release" after three years. Log growth rates are measured on an annualized scale. NBER recessions indicator in gray.

For illustrative purposes, we depict the averages (standard deviations) of the measurement errors $\Delta GDP_t^k - \Delta GDP_t^{12}$ in red (blue) for $k = 1, \ldots, 11$ in Figure 3.4. It can be seen that preliminary releases are at least unconditionally unbiased proxies for the final release data as the difference is always close to zero. Test statistics (not reported) also indicate that the null hypothesis of unconditional unbiasedness cannot be rejected for any $k = 1, \ldots, 11$. Similarly, the standard deviation of the measurement error is downward-sloping.

Figure 3.4: Average revision errors of US GDP growth and their standard deviations.



Notes: The figure reports measurement error statistics for $\Delta GDP_t^k - \Delta GDP_t^{12}$ with $k = 1, \dots, 11$, the difference between the 12th release of US GDP growth and the corresponding earlier ones. In red dots we depict the average difference in the entire sample. The triangles in blue depict the empirical standard deviations of the observed measurement errors.

3.6.4 Data for volatility forecasting

Our data compromises all 28 DJIA constituents on December 31, 2017 that traded continuously since January 03, 2000. One-minute intraday price data is obtained from QuantQuote and is aggregated to daily measures of realized variation. Prices are measured on day t at N equidistant points in time $\tau_0, \dots, \tau_N$ denoted as $p_{0,t}, \dots, p_{N,t}$. We set τ_0 to be the market opening and τ_N to be the market closing times. Accordingly, we have N intraday returns $r_{\tau,t} = 100 \cdot (\log p_{\tau,t} - \log p_{\tau-1,t})$ from which we can derive our high-frequency variation estimators. A typical trading day in our data set begins at 9:30 and ends at 16:00. Hence, most days have $N = 78$ 5-minute returns. We only consider intraday measures of volatility in this example; that is, we discard overnight returns.

The average realized variance at day t is calculated as

$$RV_t = \sum_{\tau=1}^{N} r_{\tau,t}^2.$$

The entity RV_t has been shown to converge to the actual quadratic variation of stock returns (Andersen et al., 2003) as $N \to \infty$. However, the presence of market micro-structure noise puts a lower bound on the accuracy of the estimator. The choice of calculating RV sampled on 5-minute returns is regularly seen as being a good trade-off. As a “distorted” measure of realized volatility we also calculate $\widetilde{\mathrm{RV}}$ based on prices sampled at a 15-minute frequency. Liu, Patton, and Sheppard (2015) provide an extensive empirical assessment regarding the accuracy of a wide range of realized volatility estimators.

For the competitor model in our empirical exercise, we also compute two semivariance measures introduced by Barndorff-Nielsen, Kinnebrock, and Shephard (2010). They are defined as follows,

$$RV_t^+ = \sum_{\tau=1}^{N} r_{\tau,t}^2 \mathbb{1}_{\{r_{\tau,t} \geq 0\}} \quad \text{and} \quad RV_t^- = \sum_{\tau=1}^{N} r_{\tau,t}^2 \mathbb{1}_{\{r_{\tau,t} < 0\}}.$$

The reasoning for the decomposition of RV into RV^+ and RV^- is that negative returns have a more pronounced effect on future RV than positive returns (Patton and Sheppard, 2015). This phenomenon is typically known as the "leverage" effect. The measures are calculated separately for each stock i and they are denoted by $RV_{i,t}, \widetilde{RV}_{i,t}, \widetilde{RV}_{i,t}^+$ and $\widetilde{RV}_{i,t}^-$.

References

Amado, C., A. Silvennoinen, and T. Teräsvirta. 2019. Models with multiplicative decomposition of conditional variances and correlations. In *Financial Mathematics, Volatility and Covariance Modelling*, Vol. 2, ed. by J. Chevallier, S. Goutte, D. Guerreiro, S. Saglio, and B. Sanhaji, 217–260. Routledge, UK: Milton.

Amado, C., and T. Teräsvirta. 2008. Modelling conditional and unconditional heteroskedasticity with smoothly time-varying structure. SSE/EFI Working Paper Series in Economics and Finance, No. 691. Stockholm School of Economics, The Economic Research Institute (EFI), Stockholm.

– . 2013. Modelling volatility by variance decomposition. *Journal of Econometrics* 175:142–153.

– . 2017. Specification and testing of multiplicative time-varying GARCH models with applications. *Econometric Reviews* 36:421–446.

Amaya, D., P. Christoffersen, K. Jacobs, and A. Vasquez. 2015. Does realized skewness predict the cross-section of equity returns? *Journal of Financial Economics* 118:135–167.

Amendola, A., V. Candila, and A. Scognamillo. 2017. On the influence of US monetary policy on crude oil price volatility. *Empirical Economics* 52:155–178.

Amisano, G., and R. Giacomini. 2007. Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics* 25:177–190.

Andersen, T. G., and T. Bollerslev. 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39:885–905.

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

Andersen, T. G., T. Bollerslev, and N. Meddahi. 2005. Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica* 73:279–296.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang. 2006. The cross-section of volatility and expected returns. *Journal of Finance* 61:259–299.

– . 2009. High idiosyncratic volatility and low returns: International and further U.S. evidence. *Journal of Financial Economics* 91:1–23.

Ardia, D., K. Bluteau, K. Boudt, L. Catania, and D. A. Trottier. 2019. Markov-switching GARCH models in R: The MSGARCH package. *Journal of Statistical Software* 91:1–38.

Aruoba, S. B. 2008. Data revisions are not well behaved. *Journal of Money, Credit and Banking* 40:319–340.

Asgharian, H., A. J. Hou, and F. Javed. 2013. The importance of the macroeconomic variables in forecasting stock return variance: A GARCH-MIDAS approach. *Journal of Forecasting* 32:600–612.

Asness, C., A. Frazzini, N. J. Gormsen, and L. H. Pedersen. 2020. Betting against correlation: Testing theories of the low-risk effect. *Journal of Financial Economics* 135:629–652.

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

Baker, M., B. Bradley, and J. Wurgler. 2011. Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly. *Financial Analysts Journal* 67:40–54.

Baker, S. R., N. Bloom, and S. J. Davis. 2016. Measuring economic policy uncertainty. *The Quarterly Journal of Economics* 131:1593–1636.

Bali, T. G., R. F. Engle, and S. Murray. 2016. Empirical Asset Pricing: The Cross Section of Stock Returns. New Jersey: John Wiley & Sons.

Bali, T. G., N. Cakici, and R. F. Whitelaw. 2011. Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics* 99:427–446.

Bandi, F. M., J. R. Russell, and Y. Zhu. 2008. Using high-frequency data in dynamic portfolio choice. *Econometric Reviews* 27:163–198.

Baran, S., and S. Lerch. 2015. Log-normal distribution based EMOS models for probabilistic wind speed forecasting. *Quarterly Journal of the Royal Meteorological Society* 141:2289–2299.

Barberis, N., and M. Huang. 2008. Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review* 98:2066–2100.

Barndorff-Nielsen, O. E., S. Kinnebrock, and N. Shephard. 2010. Measuring downside risk—realized semivariance. *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*: 1–22.

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

Basu, A., I. R. Harris, N. L. Hjort, and M. C. Jones. 1998. Robust and efficient estimation by minimising a density power divergence. *Biometrika* 85:549–559.

Bates, J. M., and C. W. J. Granger. 1969. The combination of forecasts. *Operational Research Quarterly* 20:391–410.

Bekaert, G., and M. Hoerova. 2014. The VIX, the variance premium and stock market volatility. *Journal of Econometrics* 183:181–190.

Billingsley, P. 1995. Probability and Measure. 3rd ed. New York: John Wiley & Sons.

Black, F. 1972. Capital market equilibrium with restricted borrowing. *The Journal of Business* 45:444–455.

Black, F., M. C. Jensen, and M. Scholes. 1972. The capital asset pricing model: Some empirical tests. *Studies in the Theory of Capital Markets* 81:79–121.

Blitz, D. C., and P. van Vliet. 2007. The volatility effect. *The Journal of Portfolio Management* 34:102–113.

Blitz, D. C., P. van Vliet, and G. Baltussen. 2019. The volatility effect revisited. *The Journal of Portfolio Management* 46:45–63.

Bollerslev, T. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31:307–327.

Bollerslev, T., B. Hood, J. Huss, and L. H. Pedersen. 2018. Risk everywhere: Modeling and managing volatility. *Review of Financial Studies* 31:2730–2773.

Bollerslev, T., S. Z. Li, and B. Zhao. 2019. Good volatility, bad volatility, and the cross section of stock returns. *Journal of Financial and Quantitative Analysis* 55:751–781.

Bollerslev, T., A. J. Patton, and R. Quaedvlieg. 2020. Realized semibetas: Signs of things to come. *Working Paper.*

Borup, D., and J. S. Jakobsen. 2019. Capturing volatility persistence: a dynamically complete realized EGARCH-MIDAS model. *Quantitative Finance* 19:1839–1855.

Boudt, K., G. Nguyen, and B. Peeters. 2015. The low-risk anomaly revisited on high-frequency data. In *Handbook of High Frequency Trading*, ed. by G. N. Gregoriou, 397–424. Amsterdam: Elsevier.

Brier, G. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78:1–3.

Buehler, R. 1971. Measuring information and uncertainty. In *Foundations of Statistical Inference*, ed. by V. Godambe and D. Sprott, 330–341. Holt, Reinhart, & Winston, Toronto.

Caldeira, J. F., G. V. Moura, F. J. Nogales, and A. A. Santos. 2017. Combining multivariate volatility forecasts: An economic-based approach. *Journal of Financial Econometrics* 15:247–285.

Carhart, M. M. 1997. On persistence in mutual fund performance. *The Journal of Finance* 52:57–82.

Catania, L., and T. Proietti. 2020. Forecasting volatility with time-varying leverage and volatility of volatility effects. *International Journal of Forecasting*: forthcoming.

Cederburg, S., M. S. O. Doherty, F. Wang, and X. Sterling. 2020. On the performance of volatility-managed portfolios. *Journal of Financial Economics*: forthcoming.

Clarke, R., H. de Silva, and S. Thorley. 2006. Minimum-variance portfolios in the U.S. equity market. *Journal of Portfolio Management* 33:10–24.

Clements, M. P., and A. B. Galvao. 2018. Data revisions and real-time probabilistic forecasting of macroeconomic variables. *Working Paper.*

Conrad, C., A. Custovic, and E. Ghysels. 2018. Long- and short-term cryptocurrency volatility components: A GARCH-MIDAS analysis. *Journal of Risk and Financial Management* 11:23.

Conrad, C., and O. Kleen. 2020. Two are better than one : Volatility forecasting using multiplicative component GARCH-MIDAS models. *Journal of Applied Econometrics* 35:19–45.

Conrad, C., and K. Loch. 2015. Anticipating long-term stock market volatility. *Journal of Applied Econometrics* 30:1090–1114.

Conrad, C., K. Loch, and D. Rittler. 2014. On the macroeconomic determinants of long-term volatilities and correlations in U.S. Stock and crude oil markets. *Journal of Empirical Finance* 29:26–40.

Conrad, C., and M. Schienle. 2020. Testing for an omitted multiplicative long-term component in GARCH models. *Journal of Business and Economic Statistics* 38:229–242.

Corradi, V., W. Distaso, and N. R. Swanson. 2009. Predictive density estimators for daily volatility based on the use of realized measures. *Journal of Econometrics* 150:119–138.

– . 2011. Predictive inference for integrated volatility. *Journal of the American Statistical Association* 106:1496–1512.

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

Corsi, F., S. Mittnik, C. Pigorsch, and U. Pigorsch. 2008. The volatility of realized volatility. *Econometric Reviews* 27:46–78.

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

Croushore, D. 2011. Frontiers of real-time data analysis. *Journal of Economic Literature* 49:72–100.

Dawid, A. P., M. Musio, and L. Ventura. 2016. Minimum scoring rule inference. *Scandinavian Journal of Statistics* 43:123–138.

Dawid, A. P., and P. Sebastiani. 1999. Coherent dispersion criteria for optimal experimental design. *Annals of Statistics* 27:65–81.

De Pooter, M., M. Martens, and D. Van Dijk. 2008. Predicting the daily covariance matrix for S&P 100 stocks using intraday data - But which frequency to use? *Econometric Reviews* 27:199–229.

DeMiguel, V., L. Garlappi, and R. Uppal. 2009. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies* 22:1915–1953.

Diebold, F. X., and R. S. Mariano. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13:253–263.

Diks, C., V. Panchenko, and D. van Dijk. 2011. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics* 163:215–230.

Ding, Z., and C. W. Granger. 1996. Modeling volatility persistence of speculative returns: A new approach. *Journal of Econometrics* 73:185–215.

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

Dominicy, Y., and H. Vander Elst. 2015. Macro-driven VaR forecasts: From very high to very low-frequency data. Available at SSRN: https://ssrn.com/abstract=2701256.

Dorion, C. 2016. Option valuation with macro-finance variables. *Journal of Financial and Quantitative Analysis* 51:1359–1389.

Ehm, W., T. Gneiting, A. Jordan, and F. Krüger. 2016. Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 78:505–562.

Elliott, G., and A. Timmermann. 2008. Economic forecasting. *Journal of Economic Literature* 46:3–56.

Engle, R. F., and J. G. Rangel. 2008. The Spline-GARCH model for low-frequency volatility and its global macroeconomic causes. *Review of Financial Studies* 21:1187–1222.

Engle, R. F., and G. M. Gallo. 2006. A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics* 131:3–27.

Engle, R. F., E. Ghysels, and B. Sohn. 2013. Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics* 95:776–797.

Engle, R. F., and G. Lee. 1999. A long-run and short-run component model of stock return volatility. In *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive WJ Granger*, ed. by R. F. Engle and H. White, 475–497. Oxford: Oxford University Press.

Engle, R. F., V. K. Ng, and M. Rothschild. 1990. Asset pricing with a Factor-ARCH covariance structure. *Journal of Econometrics* 45:213–237.

Fama, E. F., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3–56.

– . 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116:1–22.

– . 2016. Dissecting anomalies with a five-factor model. *Review of Financial Studies* 29:69–103.

Faust, J., J. H. Rogers, and J. H. Wright. 2005. News and noise in G-7 GDP announcements. *Journal of Money, Credit and Banking* 37:403–419.

Ferro, C. A. 2017. Measuring forecast performance in the presence of observation error. *Quarterly Journal of the Royal Meteorological Society* 143:2665–2676.

Fleming, J., C. Kirby, and B. Ostdiek. 2001. The economic value of volatility timing. *Journal of Finance* 56:329–352.

– . 2003. The economic value of volatility timing using realized volatility. *Journal of Financial Economics* 67:473–509.

Frazzini, A., and L. H. Pedersen. 2014. Betting against beta. *Journal of Financial Economics* 111:1–25.

Fu, F. 2009. Idiosyncratic risk and the cross-section of expected stock returns. *Journal of Financial Economics* 91:24–37.

Ghalanos, A. 2018. rugarch: Univariate GARCH models. R package. Available on CRAN: https://cran.r-project.org/package=rugarch.

Ghysels, E., A. Plazzi, R. Valkanov, A. R. Serrano, and A. Dossani. 2019. Direct versus iterated multi-period volatility forecasts: Why MIDAS is king. *Annual Review of Financial Economics* 11:173–195.

Ghysels, E., P. Santa-Clara, and R. Valkanov. 2004. The MIDAS touch: Mixed data sampling regression models. *CIRANO Working Papers.*

– . 2005. There is a risk-return trade-off after all. *Journal of Financial Economics* 76:509–548.

– . 2006. Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics* 131:59–95.

Giacomini, R., and H. White. 2006. Tests of conditional predictive ability. *Econometrica* 74:1545–1578.

Glosten, L. R., R. Jagannathan, and D. E. Runkle. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48:1779–1801.

Gneiting, T. 2011. Making and evaluating point forecasts. *Journal of the American Statistical Association* 106:746–762.

Gneiting, T., and A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102:359–378.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review* 133:1098–1118.

Gneiting, T., and R. Ranjan. 2011. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics* 29:411–422.

Good, I. J. 1952. Rational decisions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 14:107–114.

Good, I. J. 1971. Comment on paper by Buehler. In *Foundations of Statistical Inference*, ed. by V. Godambe and D. Sprott, 337–339. Holt, Reinhart, & Winston, Toronto.

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33:2223–2273.

Haas, M., S. Mittnik, and M. S. Paolella. 2004. A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics* 2:493–530.

Hampel, F. R. 1968. Contributions to the Theory of Robust Estimation. PhD thesis, University of California, Berkeley.

– . 1971. A general qualitative definition of robustness. *The Annals of Mathematical Statistics* 42:1887–1896.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. Robust Statistics. New York: John Wiley & Sons.

Han, H., and D. Kristensen. 2015. Semiparametric multiplicative GARCH-X model: Adopting economic variables to explain volatility. Working Paper.

Han, H., and J. Y. Park. 2014. GARCH with omitted persistent covariate. *Economics Letters* 124:248–254.

Han, H. 2015. Asymptotic properties of GARCH-X processes. *Journal of Financial Econometrics* 13:188–221.

Hansen, P. R., and A. Lunde. 2006. Consistent ranking of volatility models. *Journal of Econometrics* 131:97–121.

– . 2014. Estimating the persistence and the autocorrelation function of a time series that is measured with error, 30:60–93.

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

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

Haugen, R. A., and A. J. Heins. 1972. On the evidence supporting the existence of risk premiums in the capital market. *Wisconsin Working Paper.*

– . 1975. Risk and the rate of return on financial assets: Some old wine in new bottles. *The Journal of Financial and Quantitative Analysis* 10:775–784.

Hautsch, N., L. M. Kyj, and P. Malec. 2015. Do high-frequency data improve high-dimensional portfolio allocations? *Journal of Applied Econometrics* 30:263–290.

Heber, G., A. Lunde, N. Shephard, and K. Sheppard. 2009. Oxford-Man Institute's Realized Library. Oxford, UK: Oxford-Man Institute, University of Oxford.

Hillebrand, E. 2005. Neglecting parameter changes in GARCH models. *Journal of Econometrics* 129:121–138.

Huber, P. J. 1984. Finite sample breakdown of M- and P-estimators. *The Annals of Statistics* 12:119–126.

Hyvärinen, A. 2005. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research* 6:695–708.

Jacobs, J. P., and S. Van Norden. 2011. Modeling data revisions: Measurement error and dynamics of 'True' Values. *Journal of Econometrics* 161:101–109.

Jurado, K., S. E. Ludvigson, and S. Ng. 2015. Measuring uncertainty. *American Economic Review* 105:1177–1216.

Kanamori, T., and H. Fujisawa. 2015. Robust estimation under heavy contamination using unnormalized models. *Biometrika* 102:559–572.

Karanasos, M. 1999. The second moment and the autocovariance function of the squared errors of the GARCH model. *Journal of Econometrics* 90:63–76.

Kleen, O. 2017. alfred: Downloading time series from ALFRED database for various vintages. R package. Available on CRAN: https://cran.r-project.org/package=alfred.

– . 2018. mfGARCH: Mixed-frequency GARCH models. R package. Available on CRAN: https://cran.r-project.org/package=mfGARCH.

Laurent, S., J. V. K. Rombouts, and F. Violante. 2013. On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics* 173:1–10.

Lindblad, A. 2017. Sentiment indicators and macroeconomic data as drivers for low-frequency stock market volatility. Helsinki Center of Economic Research, Discussion Paper No. 413.

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

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

Matheson, J. E., and R. L. Winkler. 1976. Scoring rules for continuous probability distributions. *Management Science* 22:1087–1096.

Mincer, J., and V. Zarnowitz. 1969. The evaluation of economic forecasts. In *Economic Forecasts and Expectations*, ed. by J. Mincer, 3–46. New York, NY: National Bureau of Economic Research.

Moreira, A., and T. Muir. 2017. Volatility-managed portfolios. *Journal of Finance* 72:1611–1644.

– . 2019. Should long-term investors time volatility? *Journal of Financial Economics* 131:507–527.

Naveau, P., and J. Bessac. 2018. Forecast evaluation with imperfect observations and imperfect models. *Working Paper.*

Nelson, D. B. 1990. ARCH models as diffusion approximations. *Journal of Econometrics* 45:7–38.

– . 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59:347–370.

Newey, W. K., and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55:703–708.

Nolte, I., and Q. Xu. 2015. The economic value of volatility timing with realized jumps. *Journal of Empirical Finance* 34:45–59.

Nonejad, N. 2017. Modeling and forecasting aggregate stock market volatility in unstable environments using mixture innovation regressions. *Journal of Forecasting* 36:718–740.

Novy-Marx, R., and M. Velikov. 2016. A taxonomy of anomalies and their trading costs. *Review of Financial Studies* 29:104–147.

Opschoor, A., D. van Dijk, and M. van der Wel. 2014. Predicting volatility and correlations with financial conditions indexes. *Journal of Empirical Finance* 29:435–447.

Ovcharov, E. Y. 2017. Proper scoring rules and Bregman divergence. *Bernoulli* 24:53–79.

Pan, Z., Y. Wang, C. Wu, and L. Yin. 2017. Oil price volatility and macroeconomic fundamentals: A regime switching GARCH-MIDAS model. *Journal of Empirical Finance* 43:130–142.

Patton, A. J. 2006. Volatility forecast comparison using imperfect volatility proxies. Quantitative Finance Research Centre, University of Technology Sydney, Research Paper 175.

– . 2011. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160:246–256.

Patton, A. J., and K. Sheppard. 2009. Evaluating volatility and correlation forecasts. In *Handbook of Financial Time Series*, ed. by T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen, 801–838. Berlin: Springer.

– . 2015. Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics* 97:683–697.

Patton, A. J., J. F. Ziegel, and R. Chen. 2019. Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics* 211:388–413.

Paye, B. S. 2012. 'Déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics* 106:527–546.

Shephard, N., and K. Sheppard. 2010. Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics* 25:197–231.

Shumway, T. 1997. The delisting bias in CRSP data. *The Journal of Finance* 52:327.

Sims, C. A. 1980. Macroeconomics and reality. *Econometrica* 48:1–48.

Timmermann, A. 2006. Forecast combinations. In *Handbook of Economic Forecasting*, ed. by G. Elliott, C. W. J. Granger, and A. Timmermann, 1:135–196. Amsterdam: Elsevier.

Wallis, K. F. 2004. An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review* 189:64–71.

– . 2014. The two-piece normal, binormal, or double Gaussian distribution: Its origin and rediscoveries. *Statistical Science* 29:106–112.

Wang, F., and E. Ghysels. 2015. Econometric analysis of volatility component models. *Econometric Theory* 31:362–393.

Wendel. 1948. Note on the gamma function. *The American Mathematical Monthly* 55:563–564.