Bespoke Realized Volatility: Tailored Measures of Risk for Volatility Prediction

This version: December 30, 2022

Andrew J. Patton[a,*], Haozhe Zhang[a]

[a] *Department of Economics, Duke University*

Abstract

Standard realized volatility (RV) measures estimate the latent volatility of an asset price using high frequency data with no reference to how or where the estimate will subsequently be used. This paper presents methods for "tailoring" the estimate of volatility to the application in which it will be used. For example, if the volatility measure will be used in a specific parametric forecasting model, it may be possible to exploit that information and construct a better measure of volatility. We use methods from machine learning to estimate optimal "bespoke" RVs for heterogeneous autoregressive (HAR) and GARCH-X forecasting applications. We apply the methods to 886 U.S. stock returns and find that bespoke RVs significantly improve out-of-sample forecast performance. We find that the bespoke RV places more weight on data from the end of the trade day, and that the resulting volatility forecasts are more responsive to news than benchmark forecasts.

Keywords: Volatility forecasting, Machine learning, High frequency data
JEL: C22, C51, C53, C58

[*]For helpful comments we thank Tim Bollerslev, Anna Bykhovskaya, Adam Clements, Peter Hansen, Ole Maneesoonthorn, Jia Li, Michael Smith, Jun Yu and seminar participants at Duke, Singapore Management University, Melbourne University and the University of Sydney Time Series and Forecasting Symposium. The first author also thanks UNSW Business School, where part of this work was completed, for their hospitality.

[*]Corresponding author: Department of Economics, Duke University, 213 Social Sciences Building, Durham NC 27708-0097, USA. `andrew.patton@duke.edu`

1. Introduction

Discussing the empirical success of models based on high-frequency measures of volatility, Andersen, Bollerslev, Diebold and Labys (2003) make the point that *"[t]he essence of forecasting is quantification of the mapping from the past and present into the future. Hence, quite generally, superior estimates of present conditions translate into superior forecasts of the future."* In the subsequent two decades a large literature on high frequency financial econometrics has emerged, containing many studies confirming that using more accurate measures of volatility in a forecasting model indeed leads to better predictions.

Advances in the financial econometrics literature in the last two decades have produced to measures of volatility that are more efficient (e.g. Ait-Sahalia, Mykland and Zhang, 2005; Bandi and Russell, 2008), robust to micro-structure noise (e.g. Zhang, Mykland and Aït-Sahalia, 2005; Barndorff-Nielsen, Hansen, Lunde and Shephard, 2008; Jacod, Li, Mykland, Podolskij and Vetter, 2009), and robust to jumps in the price process (e.g. Barndorff-Nielsen and Shephard, 2004; Mancini, 2009; Andersen, Dobrev and Schaumburg, 2012). These advances share the feature that they seek an improved measure of volatility for later use in a variety of unknown applications. We consider the construction of a volatility measure from a different perspective: Can a volatility measure be improved by tailoring it for its eventual use in a volatility forecasting model?

The distinction between an "all purpose" estimator of volatility and one that is tailored to a specific application mimics the distinction between supervised and unsupervised machine learning algorithms. In unsupervised learning, data are analyzed without the use of an outcome measure. This is analogous to the above-mentioned volatility measures, where we seek a good (somehow defined) measure of volatility that works in a variety of unknown future applications. In supervised learning the algorithm is tailored to the problem at hand. This paper focuses on the case that we know that the resulting measure will be used in a specific forecasting model, to predict a specific asset's volatility. We exploit that information to obtain a "bespoke" measure of volatility for that application. In so doing, we may obtain a worse general-purpose estimator of volatility, but it is hoped that it is a better measure of volatility for the specific purpose of volatility prediction.

To make this idea concrete, consider the widely-used heterogeneous autoregressive (HAR) model of Corsi (2009):

$$
\begin{aligned}
RV_t &= \beta_0 + \beta_d RV_{t-1} + \beta_w \frac{1}{4}\sum_{j=2}^{5} RV_{t-j} + \beta_m \frac{1}{16}\sum_{j=6}^{21} RV_{t-j} + e_t && (1)\\
\text{where} \quad RV_{t-j} &\equiv \sum_{i=1}^{M} r_{i,t-j}^2 && (2)
\end{aligned}
$$

and $r_{i,t-j}$ is the i^{th} high frequency return on day $t-j$. In standard applications, RV_t is constructed as the sum of squared five-minute returns over day t, as in equation (2), which for stocks on the New York Stock Exchange means the sum of $M = 78$ such returns. Realized volatility can be shown to be consistent as the sampling interval shrinks to zero (Jacod, 2018), it is fully efficient under some regularity conditions (Jacod and Protter, 1998; Jacod, 2008), and works well in a variety of empirical applications (Liu, Patton and Sheppard, 2015). We study whether we can obtain better forecasts by altering the construction of realized volatility from that in equation (2) to exploit the knowledge that it will subsequently be used in the model in equation (1).[1]

We exploit recent advances in the estimation of deep neural networks (DNNs) to flexibly construct "bespoke" measures of volatility for use in a forecasting model. We find that being completely flexible in the construction leads to poor out-of-sample forecast performance, even when the tuning parameters of the estimation algorithm are carefully selected. However after imposing some economically motivated structure on the tailoring we obtain forecasts that significantly outperform benchmark forecasts.

Our empirical analysis uses high frequency data on all stocks that were ever a constituent of the S&P 500 index over the period January 1995 to December 2019, a total of 886 securities. In our main analysis we take the HAR model of Corsi (2009) as the predictive model, and in Section 4 we further consider the GARCH-X model (Engle, 2002) and refinements of the HAR model, such as the "continuous HAR" of Andersen, Bollerslev and Diebold (2007) and the "semi HAR" of Patton and Sheppard (2015). In

[1] Importantly, we only consider tailoring the terms on the right-hand-side of equations like (1); we leave the target variable as standard RV, with the motivation that it is a good measure of the unknown true volatility.

all cases we find significant improvements in out-of-sample forecast performance when using RVs tailored to the application.

We investigate the sources of the predictive gains from using bespoke RVs and find two primary channels. First, we find that models using bespoke RVs place greater weight on more recent lags than those using standard RVs. As more recent data tends to be more useful for prediction, this makes models using bespoke RVs more responsive to news. Second, we find that the weights attached to intra-daily returns in the bespoke RV are different from flat (as they are for standard RV) and also different from a "time-of-day" pattern motivated by measurement error considerations. Instead, the weights are low at the start of the trade day, consistent with measurement error considerations, increase slowly until the middle of the day, and then sharply increase over the last two hours of the trade day. This pattern is consistent with an information channel: returns from the latter part of the day is closest to the returns that forms part of the target variable, and thus are particularly valuable for forecasting. We find evidence in support of this explanation via multi-step-ahead forecasts.

Our work is related to the enormous literature on using high frequency data for volatility forecasting, as reviewed in Bollerslev, Engle and Nelson (1994), Poon and Granger (2003) and Andersen, Bollerslev, Christoffersen and Diebold (2006). It is also linked to the growing literature in applying machine learning methods in econometrics (e.g. Chernozhukov, Hansen and Spindler, 2015; Mullainathan and Spiess, 2017; Athey and Imbens, 2019) and finance (Gu, Kelly and Xiu, 2020; Freyberger, Neuhierl and Weber, 2020; Bianchi, Büchner and Tamoni, 2021; Patton and Weller, 2022). Within this growing machine learning in economics literature, our work is particularly related to applications of these methods for volatility forecasting (e.g. Bucci, 2020; Filipović and Khalilzadeh, 2021; Li and Tang, 2021; Christensen, Siggaard and Veliyev, 2022; Patton and Zhang, 2022; Reisenhofer, Bayer and Hautsch, 2022). Our study of high-frequency returns for predicting lower-frequency volatility also links our analysis to the "mixed data sampling" (MIDAS) models introduced by Ghysels, Santa-Clara and Valkanov (2004), and used for volatility forecasting by Ghysels, Santa-Clara and Valkanov (2006).

The rest of the paper is structured as follows: Section 2 introduces the bespoke RVs models in detail and draws connections to standard RV estimation and the standard HAR model. Section 3 presents the results on the out-of-sample forecasting performance of the competing models when applied to 886 U.S. equities. Section 4 presents results using alternative forecasting models, demonstrating the generalizability of bespoke RVs. Section 5 concludes. A supplemental appendix contains some additional details and results.

2. Constructing bespoke realized volatilities

The standard realized variance estimator, given in equation (2), can be shown (Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2002; Andersen, Bollerslev, Diebold and Labys, 2001) to be consistent for the true latent quadratic variation of an asset price process.[2] This measure has also been found to be useful for forecasting future volatility, as it is a more accurate measure of current volatility than squared daily returns, as used in ARCH/GARCH models (Engle, 1982; Bollerslev, 1986).

A key feature of the usual RV is that it is an *equal-weighted* sum of the high frequency squared returns. In constructing our "bespoke RVs" we will relax this assumption and estimate the optimal weight attached to each high frequency return. When using five-minute returns on stocks traded on the New York Stock Exchange, we have 78 such returns to consider.

2.1. Bespoke RVs for HAR models

The HAR model can be interpreted as a autoregression of order 21 with parameter equality constraints to reduce the number of free parameters from 21 to three (plus the intercept). In the most flexible bespoke RV forecast, we relax both the equal weights in the construction of RV and the HAR parameter constraints on the AR(21) process to

[2] Extensions of standard RV that are robust to market microstructure effects and/or jumps are discussed in the introduction.

obtain:

$$\begin{aligned} RV_t &= \beta_0 + \sum_{j=1}^{21} \widetilde{RV}_{t-j}(\boldsymbol{\gamma}_j) + e_t \qquad (3) \\ \text{where} \quad \widetilde{RV}_{t-j}(\boldsymbol{\gamma}_j) &\equiv \gamma_{i,j} r_{i,t-j}^2 \qquad (4) \end{aligned}$$

This simple-looking model is very flexible, with a total of $1 + 21 \times 78 = 1,639$ free parameters. Being linear, it is possible to estimate this model via standard OLS, however with the sample sizes available in practice OLS unsurprisingly performs very poorly. Instead, we treat this model as a single-layer neural network model and estimate it using methods from machine learning. We describe the estimation method in detail in the Section 2.3.

We next consider a hybrid between the fully flexible model in equation (3) and the restrictive standard HAR: we impose the constraint that the "daily," "weekly," and "monthly" lags in the model satisfy the parameter equality constraints in the HAR structure, but we allow each of these terms to be flexible functions of the underlying high-frequency returns:

$$RV_t = \beta_0 + \widetilde{RV}_{t-1}(\boldsymbol{\gamma}_d) + \frac{1}{4}\sum_{j=2}^{5} \widetilde{RV}_{t-j}(\boldsymbol{\gamma}_w) + \frac{1}{16}\sum_{j=6}^{21} \widetilde{RV}_{t-j}(\boldsymbol{\gamma}_m) + e_t \qquad (5)$$

This model has "only" $1 + 3 \times 78 = 235$ parameters and is thus much more parsimonious than the fully flexible specification.

The final bespoke RV we consider is one that imposes some smoothness on the weights attached to the high frequency returns. Research on the empirical characteristics of intra-daily asset returns (see, e.g., Wood, McInish and Ord, 1985; Harris, 1986; Andersen and Bollerslev, 1998) shows that market conditions vary over the trade day, but generally smoothly. We impose this smoothness by using a cubic spline (see Judd, 1998, for further details on this interpolation method) for the intra-day weights:

$$RV_t = \beta_0 + \widetilde{RV}_{t-1}(f(\mathbf{c}_d)) + \frac{1}{4}\sum_{j=2}^{5} \widetilde{RV}_{t-j}(f(\mathbf{c}_w)) + \frac{1}{16}\sum_{j=6}^{21} \widetilde{RV}_{t-j}(f(\mathbf{c}_m)) + e_t \qquad (6)$$

where $f(\mathbf{c})$ returns a 78×1 vector of weights based on a cubic spline. With nodes every hour or half-hour, the parameter vector $\mathbf{c}$ is of length 7 or 13, leading to a total of either 22 or 40 free parameters.

2.2. A time-of-day adjusted RV

We consider one additional estimator, lying in between standard RV and bespoke RV, motivated by two well-known features of volatility. Firstly, volatility has a prounounced diurnal pattern, with volatility being highest at the open and close of the trade day, and lowest around the middle, see Andersen and Bollerslev (1997) and Andersen and Bollerslev (1998) for example. Figure 1 confirms this pattern for the SPY in our sample period. Secondly, the estimation error in sample variance is increasing with the level of the variance. For example, in a simple i.i.d. setting, the asymptotic variance of the sample variance is proportional to the true variance squared:

$$\sqrt{T}(\hat{\sigma}^2_T - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$$

Combining the diurnal pattern in volatility with the fact that the accuracy of volatility changes with its level suggests a simple alternative to standard RV: a "time-of-day" adjusted RV, where we use the inverse of the average level of volatility as a weight function:

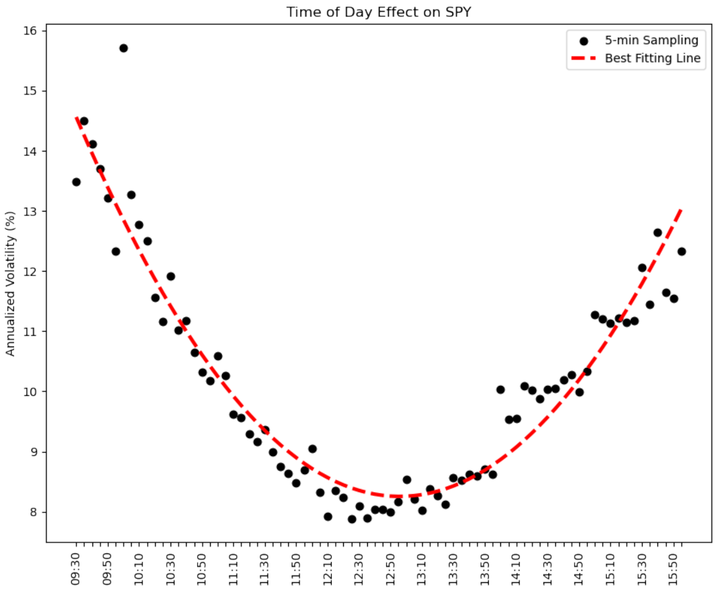
$$\begin{aligned} RV_t^{TOD} &= \sum_{i=1}^{M} \omega_i r_{i,t}^2 & (7) \\ \text{where} \quad \omega_i^{-1} &= \frac{1}{T}\sum_{t=1}^{T} r_{i,t}^2 & (8) \end{aligned}$$

To avoid contaminating our out-of-sample comparisons, we estimate the RV_t^{TOD} weights, ω_i, using only the training sample.

Using RV_t^{TOD} in the familiar HAR specification yields:

$$RV_t = \beta_0 + \beta_d RV_{t-1}^{TOD} + \beta_w \frac{1}{4}\sum_{j=2}^{5} RV_{t-j}^{TOD} + \beta_m \frac{1}{16}\sum_{j=6}^{21} RV_{t-j}^{TOD} + e_t \tag{9}$$

Figure 1: Time of Day Effect on SPY

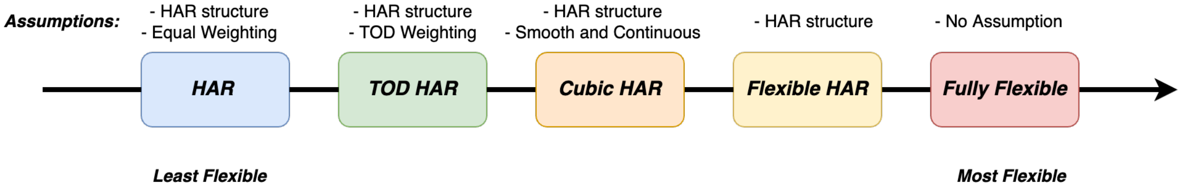


Note: This figure illustrates the time of day pattern for the SPY ETF from 1995 to 2019. The black dots are the average 5-min sample intra-day volatility (annualized) and the dashed red line is the best fitting line. We can see the clear U-shape and Diurnality pattern.

Note that RV_t^{TOD} is *not* a bespoke RV; it is not customized for a specific forecasting model, rather it is a simple alternative to equal-weighting intra-daily returns, motivated by measurement error considerations. We consider this volatility measure as another benchmark that our proposed bespoke RVs must beat in order to be considered a success.

Figure 2 summarizes the models that will be considered in our out-of-sample analysis in the next section. These range from the benchmark HAR model using standard RV as a predictor variable as the simplest specification, to the "fully flexible" model that imposes the fewest constraints.

Figure 2: Main Models Relationship



Note: This figure illustrates the relationship among the five models considered in the out-of-sample analysis.

2.3. Estimation of Bespoke RVs

The specifications in equations (3) and (5) are both linear and could easily be estimated via OLS, but this is unlikely to good out of sample performance (and we demonstrate this empirically below) as the model is severely over-parameterized. The specification in equation (6) is nonlinear, and requires a numerical optimization method. We estimate all of these models using the gradient descent approach, described below, designed for training deep neural networks. Such models are very flexible and users of these models are aware of the need for methods to tame the over-fitting problem. As is common in hte machine learning literature, we split our sample period into "training," "validation," and "testing" samples to estimate the models, select the hyperparameters, and compute out-of-sample forecasts.

2.3.1. Stochastic gradient descent methods

We adopt the mini-Batch stochastic gradient descent algorithm (Robbins and Monro, 1951; Bilmes, Asanovic, Chin and Demmel, 1997) with a gradually decreasing of learning rate. We estimate the parameters using the following algorithm:

1. Initialize the model with HAR model coefficients estimated on the training sample.[3]
2. For each combination of hyperparameters, described below, estimate the model in the training sample.
3. Evaluate predictive performance on validation sample, and select the hyperparameters that lead to the best forecasting performance.
4. Fix the hyperparameters at their optimal values, and estimate the model parameters on the entire in-sample data (training and validation samples combined).
5. Evaluate the predictions on the testing sample (the out-of-sample period).

There are a total of five hyperparameters in the optimization algorithm for the "fully flexible" and "flexible HAR" methods, and six for the "cubic HAR" method:

[3] We also considered using many random starting and then ensembling the resulting estimates, as is common in the machine learning literature, and the results are qualitatively the same.

- $LearningRate \in \{0.1, 0.01\}$. The learning rate controls the step size of each mini-Batch gradient update in the optimization algorithm.

- $BatchSize \in \{512, 2048\}$. The batch size controls how many observations we use to compute the gradient direction for each update in the optimization algorithm.

- $NumberofEpochs \in \{50, 100, 250\}$. This controls the number of times that the algorithm works through the training data set to compute the gradients for the update.

- $StepSize \in \{0.1, 0.5\}$. The step size controls how frequently we reduce the learning rate (if at all). The numbers are as a fraction of the number of epochs. For example, if the step size is 0.1 and the the number of epochs is 100, then we reduce the learning rate by gamma (described below) after every $0.1 \times 100 = 10$ epochs.

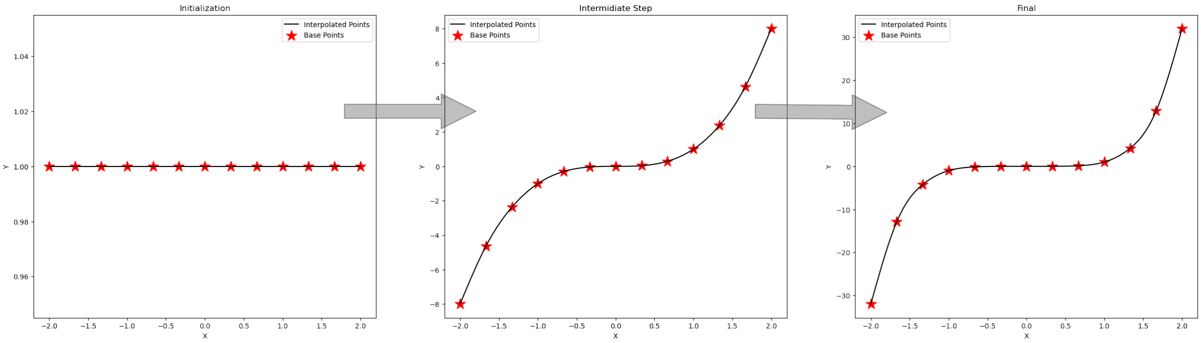
- $Gamma \in \{0.1, 0.25, 1\}$. Gamma controls how much to reduce the learning rate. For example, if the learning rate is 0.1 and gamma is 0.25, then when it is time to reduce learning rate, the updated learning rate will be $0.1 \times 0.25 = 0.025$.

- $NumNodes \in \{7, 13\}$. This parameter controls how flexible the spline function is.

We tune the hyperparameters separately for each stock in our analysis. Section S.7 in the supplemental appendix reports summary statistics on how the optimal hyperparameters vary across stocks.

In estimating the cubic spline model, we need to implement a cubic spline layer, similar to the linear and convolution layers in PyTorch, the Python package we use for estimation, where the layer can initialize values for the base points (K) and generate cubic spline interpolations for the final desired number of points (M). Note that the parameters for this cubic spline layer is the K initialized base points.and then we use the standard back-propagation and mini-Batch stochastic gradient descent framework, where we gradually find the optimal parameter values by iterating through all the batches and epochs. As an illustration, Figure 3 demonstrates how the cubic spline layer may gradually recover its optimal parameters as the mini-Batch stochastic gradient descent optimization algorithm evolves. The cubic spline layer starts out with a random initialization for the values of the interpolating nodes and then gradually converges to the optimal interpolating nodes for the training sample.

Figure 3: Illustration of Cubic Spline Layer



Note: This figure illustrates how the cubic spline layer can recover $y = x^5$ on $[-2, 2]$, with 13 base points and then interpolate to 78 points. Here the cubic spline layer parameters are the red stars and the black line represents the interpolated output. The leftmost picture represents the initialization of the layer parameters, and the middle figure represents the process of finding the optimal parameters through miniBatch gradient descent. Finally, the right most picture representing the final optimal parameter values and its interpolation output.

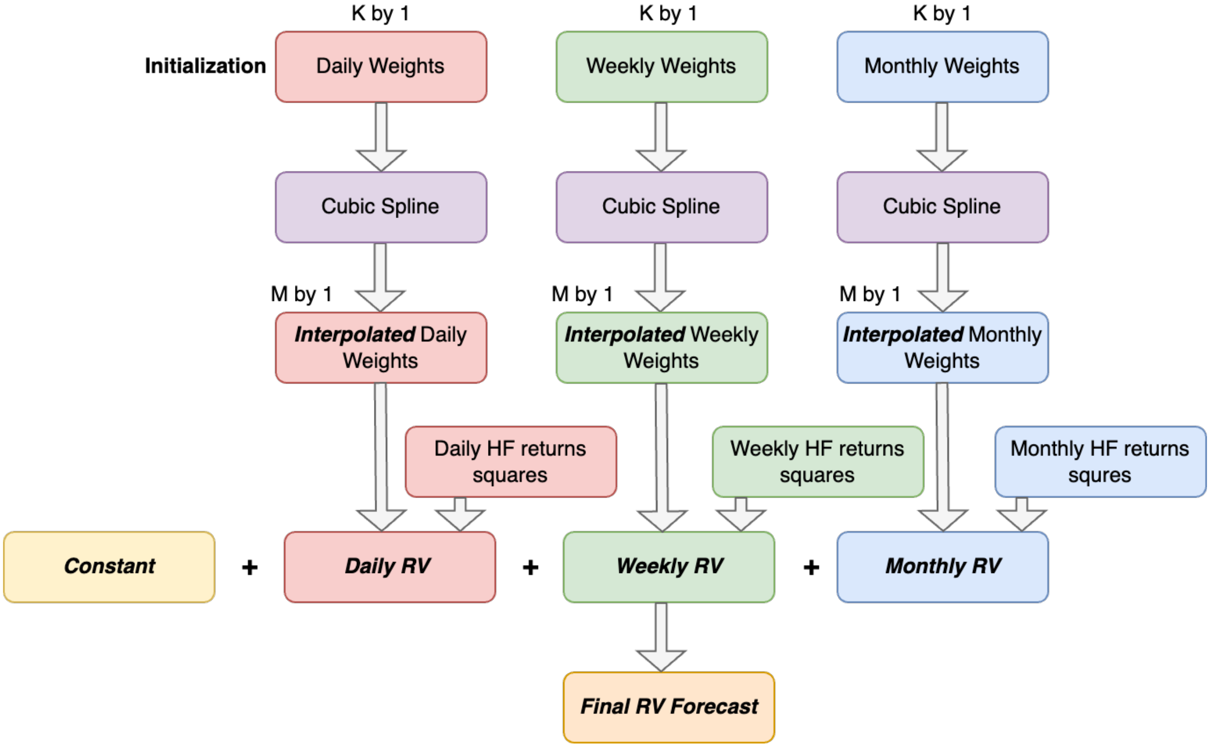
In Figure 4 we present the model architecture for the cubic spline model. (The architecture for the other two models is simliar, but without the second and third layers.) We initialize cubic spline interpolating nodes for daily, weekly and monthly lags separately, and then use the cubic spline layer to generate the initial bespoke weights. Then we combine the initial bespoke weights with the lagged high frequency returns squares and use a linear layer with a constant to construct the forecast. After this, we optimize the interpolating nodes and the constant term through the miniBatch gradient descent and reach the optimal bespoke weights for the cubic HAR model.

2.3.2. Regularized Regression Models

In addition to the stochastic gradient descent methods described above, we also consider more familiar regularization methods, applied to the "flexible HAR" model in equation (5). Denoting the usual, non-penalized, objective function as $L_T(\boldsymbol{\beta})$ we consider the penalized loss:

$$\bar{L}_T(\boldsymbol{\beta}; \alpha, \lambda) = L_T(\boldsymbol{\beta}) + \alpha(\lambda\|\boldsymbol{\beta}\|_1 + (1-\lambda)\|\boldsymbol{\beta}\|_2)$$

Figure 4: Cubic HAR Model Architecture



Note: This figure illustrates the cubic HAR model architecture. In particular, it shows how we use miniBatch gradient descent to iteratively solve the optimal parameters $(3 \times K + 1)$ and then use them to get the cubic spline interpolated weights for constructing final RV forecast. Here K represents the number of basis points for cubic spline interpolation or the true number of parameters, and M represents the desired number of points as cubic spline interpolation output.

This formulation allows us to nest four standard shrinkage methods: No penalty ($\alpha = 0$), ridge regression ($\lambda = 0,\ \alpha \geq 0$), LASSO ($\lambda = 1,\ \alpha \geq 0$), and elastic net ($\lambda \in [0, 1],\ \alpha \geq 0$). We consider the following values for these hyperparameters:

$$\begin{aligned} \alpha &\in 0, 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 2.5, 5, 10, 50, 100, 1000 \\ \lambda &\in 0, 0.25, 0.5, 0.75, 1 \end{aligned}$$

3. Out-of-sample forecast performance of bespoke realized volatilities

3.1. Data description

Our empirical analysis is based on high frequency stock prices from the Trades and Quotes (TAQ) database, spanning the period from January 1995 to December 2019, a total of 6,293 days. We include every stock that was ever a constituent of the S&P 500 index during this period, and we follow Barndorff-Nielsen, Hansen, Lunde and Shephard (2009) for the data cleaning process, retaining only stocks with at least 2,000 observations in the sample period. This yields a total of 886 different stocks for our analysis. Following Liu, Patton and Sheppard (2015), we use five-minute sampling for our high frequency returns throughout.

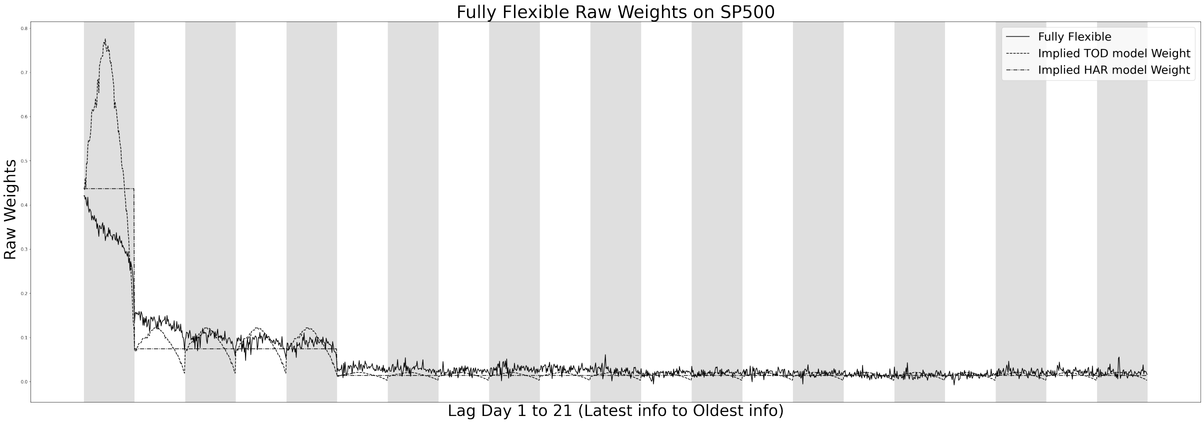
We follow common practice in the machine learning literature (see, e.g., Christensen, Siggaard and Veliyev, 2022) and split the available sample period for a given stock into a training sample (first 60% of data), a validation sample for choosing hyperparameters (next 20%), and a test sample for out-of-sample comparisons (final 20%). The first 80% of the sample is the full in-sample period. For models that do not involve any hyperparameters search we simply estimate the models on the in-sample period and evaluate on the test sample. For models that involve hyperparameters, we estimate the models on the training sample, and select the hyperparameters based on the validation sample performance. We then we re-estimate the models using the optimal hyperparameters on the full in-sample period, and evaluate on the out-of-sample period.

3.2. Optimal weights for bespoke RVs

This section presents the optimal bespoke RV weights across the three degrees of tailoring that we consider (fully flexible, flexible HAR, and cubic HAR). In Figure 5 we present the bespoke weights implied by the "fully flexible" model, averaged across all 886 stocks in our sample, and for comparison we also present the weights implied by the standard HAR (which appear as a step function) and the TOD-HAR (which appear as an inverse-U layered on a step function).

The weights for the fully-flexible bespoke RV are similar to the weights for standard RV for daily lags 7 to 21, while they are lower for the first daily lag, and slightly above for lags 2 through 6. For all lags, the estimated weights appear to have non-negligible estimation error, which is unsurprising given that each of these 1,638 weights are freely estimated. Foreshadowing our forecast comparison results, the estimation error in these weights make the fully flexible model perform significantly worse than the benchmark HAR forecasts out of sample.

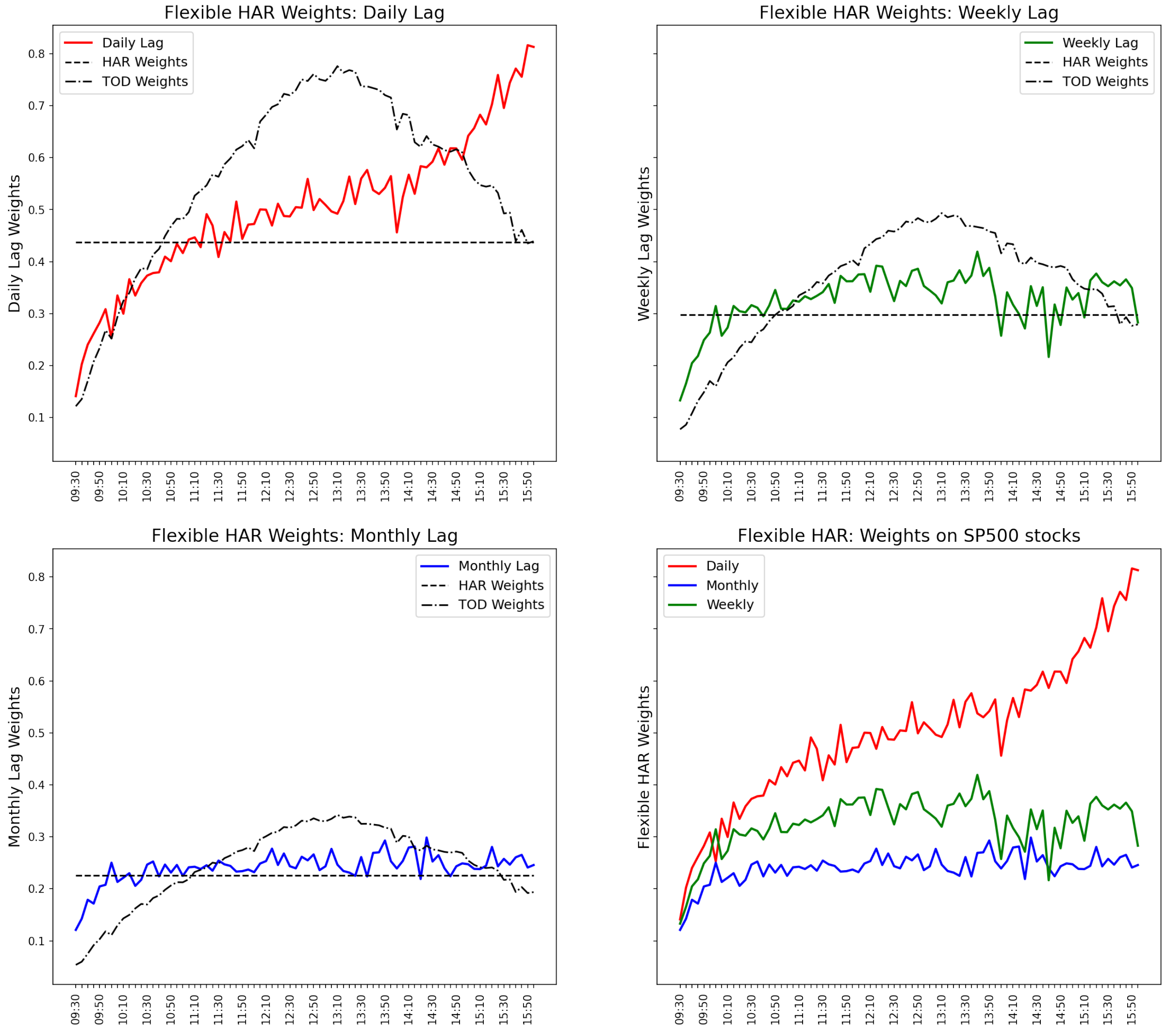
Figure 5: Cross Sectional Average Optimal Fully Flexible Weights: S&P500



Note: This figure depicts the cross-sectional average optimal weights implied by the "fully flexible" model, along with the equal-weighting scheme and the time-of-day weighting scheme. The x-axis runs from the most-recent high-frequency return to least recent, and each of the 21 days is marked by a gray or white region.

Figure 6 presents the optimal bespoke weights for the "flexible HAR" model, which imposes that RVs lagged 2 through 5 periods share the same "weekly" weight function, and the RVs lagged 6 through 21 share the same "monthly" weight function, while the first lag gets its own "daily" weight function. We find that the daily lag weights, despite

Figure 6: Cross Sectional Average Optimal Flexible HAR Weights: S&P500

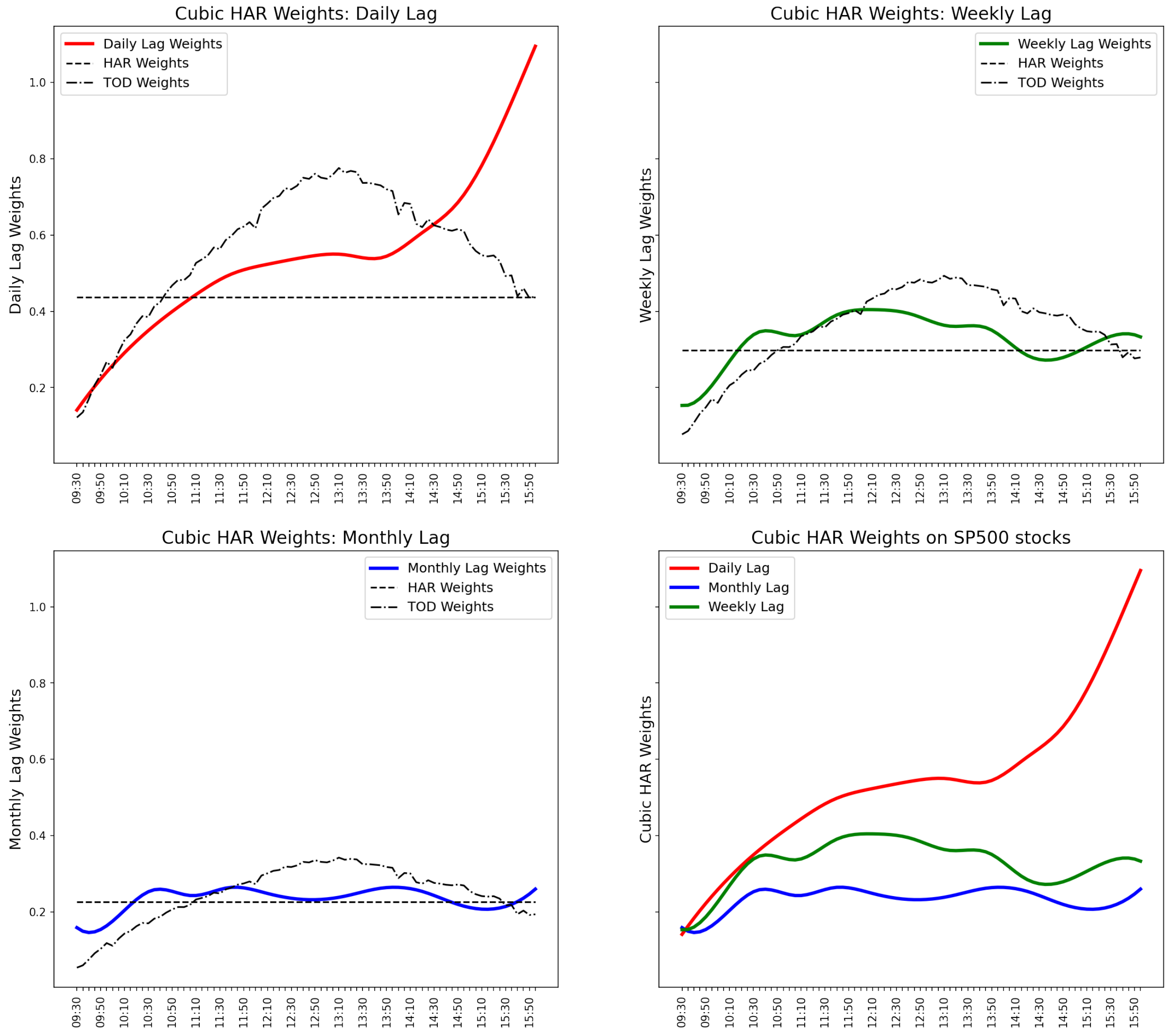


Note: This figure depicts the cross sectional average optimal weights implied by Flexible HAR model, along with its comparisons with the equal-weighting scheme and the time-of-day weighting scheme in the S&P500 cross section. The upper left corner depicts the daily lag weights, upper right is the weekly lag weights, the lower left is the monthly lag weights, and the lower right presents all three for ease of comparison.

still being somewhat noisy, clearly display an up-weighting the end of day information. For the weekly and monthly lags weights, we find them again being somewhat noisy, but roughly showing a flat shape with slight down-weighting at the beginning of the trade day.

Finally, we present the bespoke RV weights when we impose smoothness across the trade day using a cubic spline. Figure 7 presents the optimal bespoke weights for the daily, weekly and monthly weights in the cross section of S&P500. The most prominent feature

Figure 7: Cross Sectional Average Optimal Cubic HAR Weights: S&P500



Note: This figure depicts the cross sectional average optimal weights implied by the Cubic HAR regressions, along with its comparisons with the equal-weighting scheme and the time-of-day weighting scheme in the S&P500 cross section. The upper left corner depicts the daily lag weights, upper right is the weekly lag weights, the lower left is the monthly lag weights, and the lower right presents all three for ease of comparison.

of the optimal weights is the strong increase in weights in the daily lag towards the end of the trade day. We investigate the source of this feature in Section 3.4 below. It is also noteworthy that the cubic HAR daily lag weights roughly track the TOD weights until lunch time, but are much higher at the end of day. The weekly and monthly weights are broadly similar to each other, and hard to distinguish from either flat or TOD weights. In Section 3.4 we use out-of-sample forecast performance to determine whether these weights are indeed different from either of these benchmarks.

3.3. Comparing out-of-sample forecast performance

We now present results comparing the forecast performance of the models using bespoke RVs with the benchmark HAR model using standard RV, as well as the HAR model using the time-of-day weighted RV. For our main analysis we measure forecast accuracy using the QLIKE loss function:

$$L_{QLIKE}(RV, \widehat{RV}) = RV/\widehat{RV} - \log RV/\widehat{RV} - 1$$

We report corresponding results using the quadratic loss function, $L_{MSE}(RV, \widehat{RV}) = (RV - \widehat{RV})^2$, in the supplemental appendix.[4] We compare the predictive accuracy of competing models using Diebold-Mariano (DM) tests for each individual asset in our sample, and a panel DM test for all stocks jointly.[5]

Table 1 presents forecast comparison results comparing the baseline HAR model with various competing methods. The first four rows present comparisons with the methods presented in Figure 2, namely the TOD-HAR and three bespoke HAR models, while the bottom four rows present comparisons with different, more familiar, methods for estimating the "flexible HAR" model.

The first row of Table 1 compares the baseline HAR model with a fully flexible bespoke HAR models. We see that the baseline model outperforms the fully flexible model in 714 out of 886 individual comparisons, losing in only 172, and of those 714 "wins" 321 are statistically significant at the 5% level. Pooling the individual stocks and conducting the comparison jointly, the last column reports a panel Diebold-Mariano t-statistic of -9.3, which is strong evidence that the baseline HAR model outperforms the fully flexible model overall. These comparisons lead to the conclusion that fully flexible model is worse than the simple, familiar, and parsimonious HAR model.

The second row of Table 1 compares the baseline HAR with the "flexible HAR," which imposes the HAR structure on the lag parameters, but allows bespoke weights on each

[4] Consistent with the power analyses in Patton and Sheppard (2009), the rankings using quadratic loss are similar to those using QLIKE but the significance of the loss differences is generally weaker.

[5] For the individual DM tests we use Newey and West (1987) standard errors allowing for autocorrelation up to 10 lags. For the panel DM tests we additionally cluster by stock.

Table 1: Forecast performance of HAR vs other models for 886 S&P 500 stocks

	DM Losses		**DM Wins**		**DM t-stat**
HAR vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Fully Flexible	172	53	714	321	-9.3
Flexible HAR	677	430	209	38	4.5
Cubic HAR	731	470	155	41	21.7
TOD HAR	680	458	206	63	18.6
Ridge	445	185	441	175	-2.3
LASSO	222	42	664	299	-6.1
Elastic net	454	182	432	174	-7.2
OLS	245	46	641	272	-5.7

Note: This table reports individual and panel Diebold-Mariano tests comparing the baseline HAR model against competing models across 886 S&P 500 stocks. A positive panel DM t-statistic indicates that the competing model out-performs the HAR model, while a negative t-statstic indicates the opposite.

of the daily, weekly, and monthly lags. We see that this model performs much better than the fully flexible model: it out-performs the baseline HAR for 677 out of 886 stocks, and has a panel DM statistic of 4.5, indicating strongly significant out-performance. The third row of Table 1 imposes more structure on the bespoke RV, using a cubic spline to ensure that the bespoke weights are smooth through the trade day. We see that this improves forecast performance even further, with a panel DM statistic of 21.7. Thus imposing some economically-motivated structure on the bespoke RV leads us to a model that out-performs the baseline model by an even greater margin than that by which the fully flexible model under-performs.

Finally, the fourth row of Table 1 considers the non-bespoke, but computationally simple, TOD-HAR, where the realized variances are computing using time-of-day weights. We see that this model also significantly out-performs the baseline HAR, with a panel DM statistic of 18.6, and almost the same number of "wins" as the flexible (bespoke) HAR. These results suggest that this simple "off-the-rack" alternative RV is a significant improvement over standard equal-weighted RV for volatility forecasting.

The bottom four rows of Table 1 present alternative methods for regularizing the

Table 2: Forecast performance of Cubic HAR vs other models for 886 S&P 500 stocks

	DM Losses		**DM Wins**		**DM t-stat**
Cubic HAR vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Fully Flexible	106	39	780	571	-10.8
Flexible HAR	277	72	609	204	-12.8
TOD HAR	358	95	528	205	-2.9
HAR	155	41	731	470	-21.7
Ridge	114	32	772	434	-3.3
LASSO	48	15	838	595	-6.9
Elastic net	118	26	768	433	-10.2
OLS	54	17	832	564	-6.4

Note: This table reports individual and panel Diebold-Mariano tests comparing the cubic HAR model against competing models across 886 S&P 500 stocks. A positive panel DM t-statistic indicates that the competing model out-performs the cubic HAR model, while a negative t-statstic indicates the opposite.

flexible HAR model parameters. Recall that the flexible HAR model has a total of 235 parameters, and all of them appear linearly, meaning that familiar methods like OLS and ridge regression can be employed in place of our preferred stochastic gradient descent (SGD) optimization method. We see, however, that all of these methods lead to significantly worse performance than the baseline HAR: the panel DM t-statistics are all less than -2, while the flexible HAR model estimated using SGD significantly *out*-performs the baseline model.[6]

From the comparisons with the baseline model in Table 1, it appears that the cubic HAR is the best-performing model, but strictly those comparisons do not guarantee this interpretation is correct. Table 2 changes the reference model to the cubic HAR and compares all of the other methods to it. We see that the panel DM t-statistic is uniformly below -2, and indeed in several comparisons it is much smaller than that, confirming that the cubic HAR model is indeed the best-performing model on average.

[6] In Section S.3 of the supplemental appendix we consider shrinking the estimated OLS parameters towards the benchmark HAR parameters rather than towards zero, as is done here. We find that the optimal degree of shrinkage for this design is large, and the estimated parameters are shrunk almost all the way towards the original HAR parameters. The cubic HAR model is shown to significantly out-perform these alternative shrinkage estimators as well.

We next consider the gains from bespoke RV for multi-step ahead volatility forecasting. Table 3 presents results for forecast horizons ranging from one day to 60 days. In Panel A we see that using bespoke RV significantly improves forecast accuracy for all horizons, with the panel DM t-statistic less than -5 in all cases. The improvement is generally declining with the horizon, with the t-statistics decreasing in magnitude, and the proportion of "wins" in individual comparisons decreasing as well. Nevertheless, these results represent strong evidence that "bespoke" RVs are preferred to standard RVs, when employed in the HAR model, even for relatively long forecast horizons.

In Panel B of Table 3 we compare the bespoke RV to the simple TOD-RV. Table 1 revealed that TOD-RV is significantly better, on average, than standard RV and so this is a much tougher competitor. We find statistically significant gains from using bespoke RV in around half of the horizons considered, and in only one horizon is the ranking reversed, though in that case the difference in forecast performance is not significant. This panel thus also confirms that bespoke RVs provide important improvements in forecast performance across a range of forecast horizons.

3.4. Understanding the Optimal Bespoke Weights

In this section we seek to understand the sources of the out-performance of models using bespoke RV relative to the benchmark methods. We first investigate whether by tailoring the measure of risk to the forecasting problem at hand the model can place greater weight on the risk measure and thus react to news more quickly. To measure this, we compute the average effective regression coefficients for the daily, weekly, and monthly lagged RVs.[7] Table 4 presents the results, and shows that the bespoke RVs are more responsive than the standard RVs. The standard HAR has an average coefficient on daily lagged RV of 0.438, while the bespoke cubic HAR has an average coefficient of 0.466. In the other direction, we see that the poor-performing fully flexible HAR has an average daily coefficient of only 0.312. When the sum of the coefficients is less than one, we can interpret that sum as the weight on lagged information, and the difference

[7] We adjust the regression weights to ensure that the bespoke RVs and the standard RVs have the same unconditional mean during the in-sample period. This makes comparing coefficient magnitudes meaningful.

Table 3: Multi-day ahead volatility forecasting

	DM Losses		**DM Wins**		**DM t-stat**
Horizon (days)	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Panel A: Cubic vs HAR					
1	155	41	731	470	-21.7
2	194	31	692	366	-11.6
3	222	38	664	325	-10.5
4	247	41	639	285	-8.8
5	250	36	636	265	-18.6
20	348	81	538	248	-8.6
60	413	154	473	223	-5.4
Panel B: Cubic vs TOD HAR					
1	358	95	528	205	-2.9
2	449	120	437	146	-1.2
3	487	120	399	136	-2.6
4	496	129	390	133	1.0
5	500	131	386	124	-4.6
20	484	160	402	143	-0.3
60	458	175	428	189	-4.9

Note: This table reports individual and panel Diebold-Mariano tests comparing the cubic HAR model against the HAR (Panel A) and TOD-HAR (Panel B) models for 886 S&P 500 stocks, for various forecast horizons. A positive panel DM t-statistic indicates that the competing model out-performs the cubic HAR model, while a negative t-statstic indicates the opposite.

from one as the effective weight on the unconditional average (the intercept). This too lines up with the relative forecast performances documented in the previous subsection, and is consistent with bespoke RVs (appropriately disciplined) providing more responsive forecasts.

We next seek to understand the reason cubic HAR weights take the shape that they do. These weights were presented above in Figure 7 and present two key questions. Firstly, why do the weights on the daily lag rise in the afternoon? Secondly, are the weights on the weekly and monthly lags significantly different from either the flat weights from the standard RV or the time-of-day (TOD) weights?

We conjecture that the rising weights in the daily lagged RV come from the proximity of the afternoon to the target day. That is, returns realized in the afternoon are the

Table 4: Average regression coefficients for different models

	Daily	**Weekly**	**Monthly**	**Sum**
HAR	0.438	0.298	0.225	0.960
TOD HAR	0.465	0.295	0.204	0.964
Cubic HAR	0.466	0.294	0.217	0.978
Flexible HAR	0.442	0.298	0.224	0.963
Fully Flexible	0.312	0.384	0.346	1.041

Note: This table presents the average, across 886 S&P 500 stocks, coefficients on daily, weekly and monthly lagged RVs.

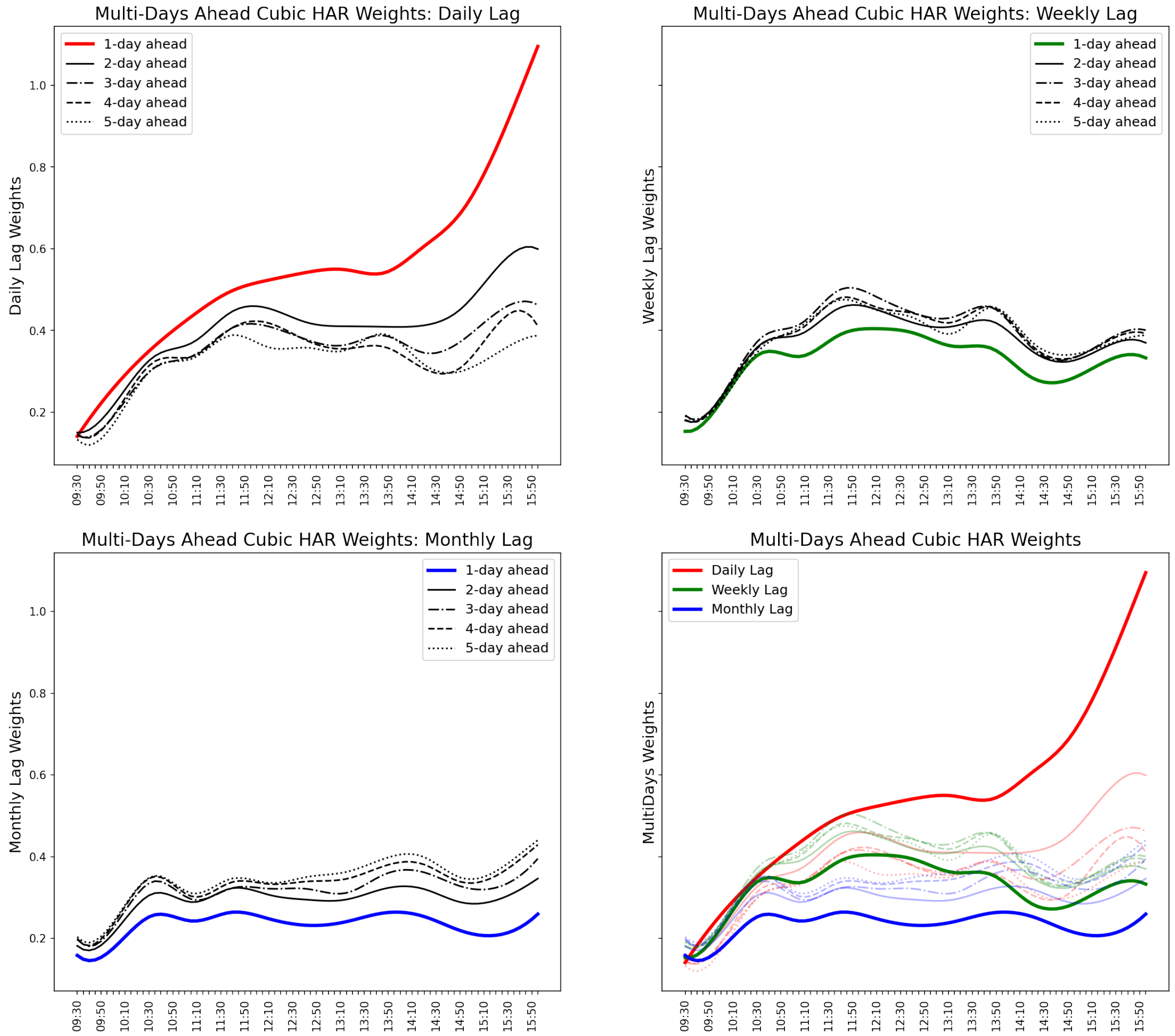
closest to, and presumably the most informative for, the returns that arise during the following day. We test this conjecture by examining the weights on the daily lag returns for longer forecast horizons. If the conjecture is correct, then the weights on the daily lagged RV should rise by less in the afternoon for longer forecast horizons. Figure 8 presents the estimated weights for forecast horizons ranging from one to five days, and we see that the weights on afternoon returns are monotonically declining as the horizon increases, consistent with this being an information effect. The weights on weekly and monthly lagged RVs increase slightly with the forecast horizon, offsetting the declining weight given to the daily lagged RVs.

Next we seek to determine whether the weights on the weekly and monthly lags significantly different from either the flat weights from the standard RV or the time-of-day (TOD) weights. We do this via an out-of-sample forecast comparison of the cubic HAR with a hybrid model that estimates the weights on the lagged daily data but imposes either equal weights or TOD weights on the lagged weekly and monthly data. We estimate these hybrid “cubic-EW” and “cubic-TOD” models using the same SGD algorithm as the original cubic HAR model.

Table 5 shows that the hybrid cubic-TOD model significantly out-performs all competing models, according to the panel DM t-statistics, aside from the cubic-EW model. The t-statistic for that comparison is 1.1, indicating no significant difference in performance on average.[8] While this analysis does not allow us to determine whether the

[8] We informally break this statistical tie by considering the number of significant differences for indi-

Figure 8: Multi-day ahead cubic HAR weights



Note: This figure depicts the cross sectional average optimal weights implied by the cubic HAR regressions for multi-days ahead forecasts in the S&P 500 cross section. Note that the upper left is the daily lag, upper right is the weekly lag, and the lower left is the monthly lag. Also the colored lines are representing weights for one day ahead forecasting.

Table 5: Forecast performance of Cubic-TOD vs other models on 886 S&P 500 stocks

	DM Losses		**DM Wins**		**DM t-stat**
Cubic-TOD vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Fully Flexible	100	35	786	567	-10.9
Flexible HAR	281	69	605	228	-14.2
Cubic	402	130	484	150	-3.7
Cubic-EW	417	123	469	136	1.1
TOD HAR	318	74	568	219	-8.1
HAR	159	37	727	471	-21.0

Note: This table reports individual and panel Diebold-Mariano tests comparing the cubic-TOD model against competing models across 886 S&P 500 stocks. A positive panel DM t-statistic indicates that the competing model out-performs the cubic-TOD model, while a negative t-statstic indicates the opposite.

optimal weights for weekly and monthly lagged information are equal or TOD shaped, Table 5 does show that it is preferable to impose either of those shapes than to try to estimate them from data.

4. Bespoke RV for alternative forecasting models

The analysis in the previous section focused on tailoring realized variance (RV) measures for application in the heterogeneous autoregressive (HAR) model of Corsi (2009). This section shows that out-of-sample forecast performance can similarly be improved when tailoring RV measures for use in alternative forecasting models. We firstly consider two refinements of the original HAR model, the “continuous” HAR (CHAR) model of Andersen, Bollerslev and Diebold (2007), which decomposes volatility into continuous and jump components, and the “semi” HAR (SHAR) model of Patton and Sheppard (2015), which decomposes volatility into upside and downside movements. We then consider a model outside the HAR family, the GARCH-X model of Engle (2002).

The CHAR model of Andersen, Bollerslev and Diebold (2007) forecasts future realized volatility using only the continuous component of volatility, discarding the component

vidual stocks: we observe that cubic-TOD significantly out-performs cubic-EW for 136 stocks, compared with 123 for cubic-EW.

coming from jumps, as that component is found to be nearly unpredictable. This model uses bi-power variation (BPV) (Barndorff-Nielsen and Shephard, 2004) to estimate the continuous component of volatility:

$$\begin{aligned} RV_t &= \beta_0 + \beta_d BPV_{t-1} + \beta_w \frac{1}{4}\sum_{j=2}^{5} BPV_{t-j} + \beta_m \frac{1}{16}\sum_{j=6}^{21} BPV_{t-j} + e_t \quad (10) \\ \text{where } \; BPV_t &= \mu_1^{-2} \sum_{i=1}^{M-1} |r_{t,i}||r_{t,i+1}| \\ \mu_1 &\equiv \sqrt{2/\pi} \end{aligned}$$

We tailor the measure of continuous volatility to the CHAR model by flexibly estimating the weights attached to each product $|r_{t,i}||r_{t,i+1}|$. Motivated by the results in Tables 2 and 5, we impose that the daily weights are smooth by using a cubic spline, and that the weekly and monthly weights use time-of-day (TOD) weights:

$$RV_t = \beta_0 + \widetilde{BPV}_{t-1}(\boldsymbol{\gamma}) + \beta_w \frac{1}{4}\sum_{j=2}^{5} BPV_{t-j}^{TOD} + \beta_m \frac{1}{16}\sum_{j=6}^{21} BPV_{t-j}^{TOD} + e_t \qquad (11)$$

where $\widetilde{BPV}_t(\boldsymbol{\gamma}) = \sum_{i=1}^{M-1} \gamma_i |r_{t,i}||r_{t,i+1}|$, and γ_i comes from a cubic spline with hourly or half-hourly nodes. BPV_t^{TOD} is defined analogously to RV_t^{TOD} in equation (7).

Next we turn to the SHAR of Patton and Sheppard (2015), which decomposes realized variance into positive and negative realized semivariances (Barndorff-Nielsen, Kinnebrock and Shephard, 2010):

$$RV_t = \beta_0 + \beta_{d,p} RV_{t-1}^{+} + \beta_{d,n} RV_{t-1}^{-} + \beta_w \frac{1}{4}\sum_{j=2}^{5} RV_{t-j} + \beta_m \frac{1}{16}\sum_{j=6}^{21} RV_{t-j} + e_t \qquad (12)$$

where $RV_t^{+} = \sum_{i=1}^{M} \max(0, r_{i,t})^2$ and $RV_t^{-} = \sum_{i=1}^{M} \min(0, r_{i,t})^2$ are the positive and negative realized semivariances. We tailor the measures of semivariance by using a cubic spline to aggregate the high-frequency positive and negative returns. As for the bespoke CHAR model, we again only estimate the weights for the daily lag, and impose TOD

Table 6: Bespoke RV for CHAR and SHAR models

	DM Losses		DM Wins		DM t-stat
Model	**Total**	**Signif**	**Total**	**Signif**	**Panel**
CHAR: Basic vs Bespoke	120	38	766	606	-19.2
SHAR: Basic vs Bespoke	164	40	722	492	-11.1

Note: This table reports individual and panel Diebold-Mariano tests comparing the CHAR and SHAR models with their bespoke counterparts, across 886 S&P 500 stocks. A positive panel DM t-statistic indicates that the original model out-performs the bespoke version, while a negative t-statstic indicates the opposite.

weights for the weekly and monthly lags:

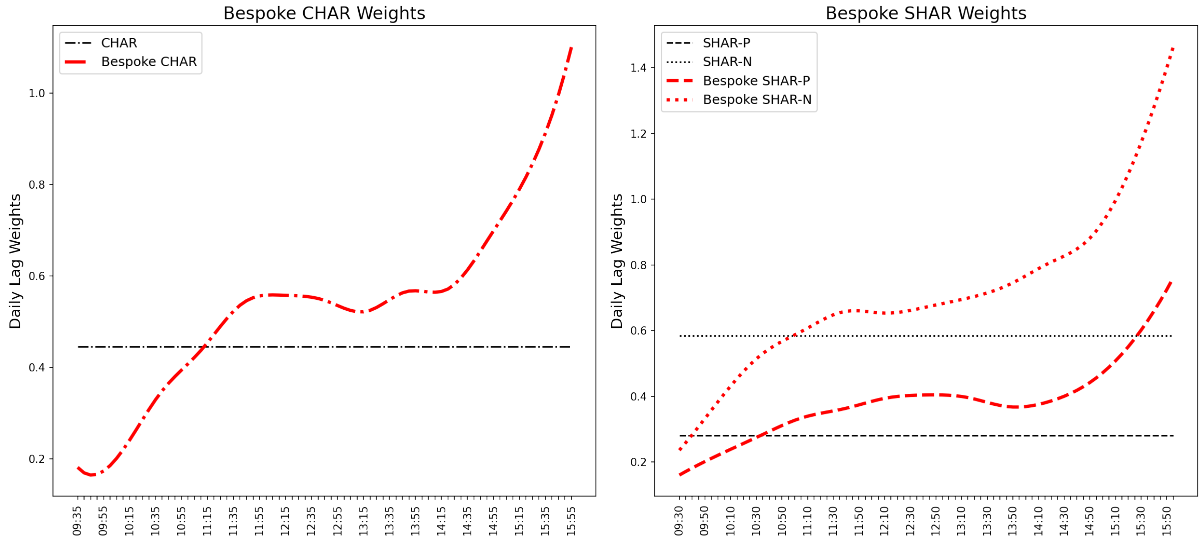
$$RV_t = \beta_0 + \widetilde{RV}^{+}_{t-1}(\boldsymbol{\gamma}_p) + \widetilde{RV}^{-}_{t-1}(\boldsymbol{\gamma}_n) + \beta_m \frac{1}{4}\sum_{j=2}^{5} RV^{TOD}_{t-j} + \beta_w \frac{1}{16}\sum_{j=6}^{21} RV^{TOD}_{t-j} + e_t \qquad (13)$$

where $\widetilde{RV}^{+}_{t}(\boldsymbol{\gamma}_p) = \sum_{i=1}^{M} \gamma_{i,p} \max(r_{t,i}, 0)^2$, $\widetilde{RV}^{-}_{t}(\boldsymbol{\gamma}_n) = \sum_{i=1}^{M} \gamma_{i,n} \min(r_{t,i}, 0)^2$, and $(\gamma_{i,p}, \gamma_{i,n})$ come from cubic splines with hourly or half-hourly nodes. RV_t^{TOD} is from equation (7).

We estimate the bespoke CHAR and SHAR models using the same methods as the models in Section 3. Table 6 presents out-of-sample forecast comparisons of the original models with their bespoke counterparts. We see that forecasts based on bespoke volatility measures significantly out-perform their benchmark alternatives. The panel DM t-statistics are less than -10 in both comparisons, and the bespoke models out-perform for over 700 of the 866 individual comparisons. These results are qualitatively as strong as the out-performance of cubic HAR over standard HAR reported in Table 1, and provide evidence that the idea of tailoring the risk measure to the forecasting model is not specific to the HAR model of Corsi (2009).

After seeing this strong forecasting performance, we are again interested in the optimal weights that lead to the improved forecasting performance. To this end, we visualize the average optimal weights from bespoke CHAR and SHAR models, and compare them with the weights from original SHAR and CHAR models, which are flat. Figure 9 shows that the optimal weights for these models both display the rise in weights in the afternoon that was observed for bespoke RV for the HAR model in Figure 7. Also, consistent with

Figure 9: Bespoke CHAR and SHAR weights



Note: This figure depicts the cross sectional average optimal weights implied by the CHAR and SHAR models along with their bespoke versions.

Patton and Sheppard (2015), we find that the weights on the negative semivariances are greater than those on the positive semivariances, both for the benchmark SHAR model and the bespoke SHAR model, indicating that negative high frequency returns are more important for forecasting one-day-ahead volatility than positive high frequency returns.

Next we explore bespoke measures of volatility for a model outside of the HAR family of models. We consider the GARCH-X model of Engle (2002), which replaces the lagged squared return in the GARCH model (Bollerslev, 1986) with the lagged realized variance.[9] Assuming a zero mean, the GARCH-X model is

$$r_t = \sqrt{h_t} z_t \tag{14}$$

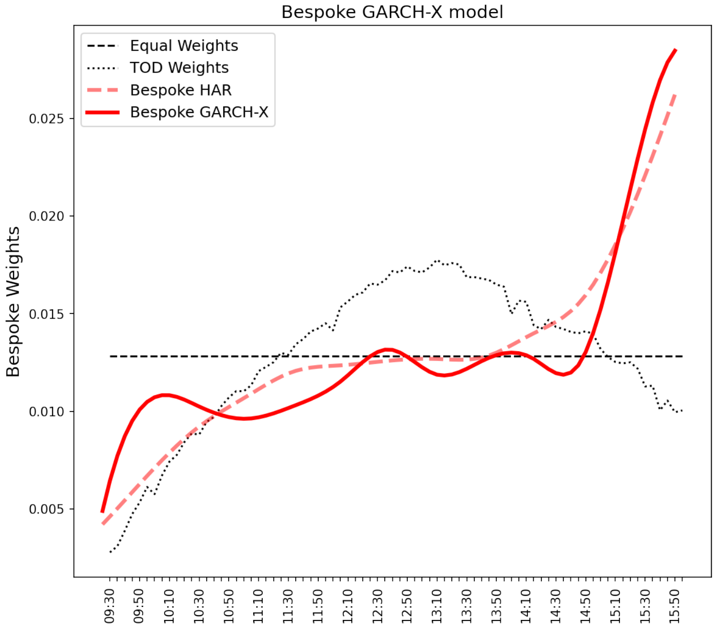
$$h_t = \omega + \beta h_{t-1} + \alpha RV_{t-1}$$

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

We construct a bespoke RV for the GARCH-X model by flexibly estimating the weights attached to the high frequency squared returns in RV_t, again using a cubic spline

[9] When focusing on one-day-ahead volatility forecasting, the HEAVY (Shephard and Sheppard, 2010) and realized GARCH (Hansen, Huang and Shek, 2012) models both reduce to a GARCH-X type model. For multi-step forecasting one of these models, or some other extension of the GARCH-X model, is required.

Figure 10: Bespoke GARCH-X weights



Note: This figure depicts the cross sectional average optimal weights implied by the Bespoke GARCH-X model weights.

to impose smoothness as a function of the time of day.

$$h_t = \omega + \beta h_{t-1} + \alpha \widetilde{RV}_t(\boldsymbol{\gamma}) \tag{15}$$

where $\widetilde{RV}_t(\boldsymbol{\gamma})$ is constructed as in the cubic HAR model in equation (6). Given the impressive performance of the simple time-of-day adjusted RV measure in the HAR analysis, we also consider a GARCH-X model with RV^{TOD} on the right-hand side in place of standard RV.

Figure 10 presents the bespoke GARCH-X model weights, as well as the bespoke HAR weights and the TOD weights for comparison. We observe that the bespoke GARCH-X weights are markedly different from both the equal weights of standard RV and the TOD weights, and are quite similar to the optimal bespoke HAR weights, starting the day low and close to the TOD weight, then rising to around the same level as the equal weighted case, and then rising strongly in the afternoon. As discussed in Section 3.4, this increase is likely driven by the additional information contained in that period for returns in the subsequent day.

Table 7: Bespoke RV for GARCH-X models

	DM Losses		DM Wins		DM t-stat
Model	**Total**	**Signif**	**Total**	**Signif**	**Panel**
GARCH-X: Basic vs Bespoke	336	68	550	186	-9.1
GARCH-X: TOD vs Bespoke	431	87	455	126	-4.3

Note: This table reports individual and panel Diebold-Mariano tests comparing the GARCH-X model with a bespoke counterpart, and with a model using RV^{TOD}, across 886 S&P 500 stocks. A negative panel DM t-statistic indicates that the bespoke model out-performs the competitor, while a positive t-statstic indicates the opposite.

Finally, we turn to forecast comparisons for this application. We evaluate the forecast performance using the QLIKE loss function and using r_t^2 as a proxy for true volatility, following the spirit of the GARCH-type models. Table 7 reports the forecast comparison results, and confirms that bespoke RVs outperform both the standard RV and RV-TOD when used in GARCH-X models. The panel DM t-statistic comparing basic and bespoke GARCH-X forecasts is -9.1, indicating very strong statistical significance of the forecast improvement. Bespoke RV also significantly outperforms TOD-RV, with a DM t-statistic of -4.3. These results confirm that tailoring the risk measure to the predictive model in which it will be used, whatever form that model takes, leads to improved out-of-sample volatility forecasts.

5. Conclusion

This paper proposes to tailor the measure of risk, such as realized variance (RV), to the specific forecasting model and the specific asset of interest in order to improve the model's forecasting performance. The resulting "bespoke RV" will not necessarily be a good "all purpose" measure of risk, but it will optimally draw on the available high frequency information to improve, if possible, the forecasting performance of the model.

We use data on all 886 stocks that were ever a constituent of the S&P 500 index over the period 1995 to 2019, and we exploit recent advances in the estimation of deep neural networks (DNNs) to flexibly construct "bespoke" measures of volatility. We find that being completely flexible in the construction of the bespoke measure leads to poor out-of-sample forecast performance, however after imposing some economically motivated structure we obtain forecasts that significantly outperform the benchmark forecasts. Our main analyses focus on bespoke RVs for the heterogeneous autoregressive (HAR) model of Corsi (2009), and as extensions we consider the GARCH-X model (Engle, 2002), the "continuous HAR" of Andersen et al. (2007), and the "semi HAR" of Patton and Sheppard (2015). In all four cases we find significant improvements in out-of-sample forecast performance when using RVs tailored to the application.

Opening the black box to understand the sources of forecast improvements from using bespoke RVs, we find two main channels. Firstly, we find that using a bespoke RV in place of a standard RV leads the model to put more weight on the risk measure, increasing the responsiveness of the forecast. Secondly, we find that the optimal bespoke RVs, across all four models, place higher weight on returns from the afternoon, which is in contrast with both the standard equal-weighted RV and an RV based on time-of-day information. We find evidence that this increased afternoon weight comes from an information channel: afternoon returns are closest to the target date, and thus more informative about the future volatility.

This paper leaves open several avenues for future work. We focused exclusively on univariate volatility models, and the important extension to multivariate models opens up questions about the optimal degree of customization potentially differing between variance and correlations, as well as the usual empirical challenges of moving to high dimension models. We focused on linear bespoke RVs, and the extension to nonlinear versions could yield further improvements, although our results suggest that imposing some structure on the bespoke RV is important for achieving forecast gains. Finally, our focus is on volatility models, but any forecasting model that uses a variable constructed from other variables or data sources may benefit from tailoring that input. We look forward to pursuing, or reading about, these ideas in the future.

References

Ait-Sahalia, Y., Mykland, P.A., Zhang, L., 2005. How often to sample a continuous-time process in the presence of market microstructure noise. Review of Financial Studies 18, 351–416.

Andersen, T.G., Bollerslev, T., 1997. Intraday periodicity and volatility persistence in financial markets. Journal of Empirical Finance 4, 115–158.

Andersen, T.G., Bollerslev, T., 1998. Deutsche mark–dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. Journal of Finance 53, 219–265.

Andersen, T.G., Bollerslev, T., Christoffersen, P.F., Diebold, F.X., 2006. Volatility and correlation forecasting. Handbook of Economic Forecasting 1, 777–878.

Andersen, T.G., Bollerslev, T., Diebold, F.X., 2007. Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. Review of Economics and Statistics 89, 701–720.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001. The distribution of realized exchange rate volatility. Journal of the American Statistical Association 96, 42–55.

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

Andersen, T.G., Dobrev, D., Schaumburg, E., 2012. Jump-robust volatility estimation using nearest neighbor truncation. Journal of Econometrics 169, 75–93.

Athey, S., Imbens, G.W., 2019. Machine learning methods that economists should know about. Annual Review of Economics 11, 685–725.

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

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2008. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. Econometrica 76, 1481–1536.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2009. Realized kernels in practice: Trades and quotes. Econometrics Journal 12, 1–32.

Barndorff-Nielsen, O.E., Kinnebrock, S., Shephard, N., 2010. Measuring downside risk: Realised semivariance, in: Bollerslev, T., Russell, J.R., Watson, M.W. (Eds.), Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle. Oxford University Press, Oxford, pp. 117–136.

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

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

Bianchi, D., Büchner, M., Tamoni, A., 2021. Bond risk premiums with machine learning. Review of Financial Studies 34, 1046–1089.

Bilmes, J., Asanovic, K., Chin, C.W., Demmel, J., 1997. Using PHiPAC to speed error back-propagation learning, in: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4153–4156.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., Engle, R.F., Nelson, D.B., 1994. ARCH models. Handbook of Econometrics 4, 2959–3038.

Bucci, A., 2020. Realized volatility forecasting with neural networks. Journal of Financial Econometrics 18, 502–531.

Chernozhukov, V., Hansen, C., Spindler, M., 2015. Valid post-selection and post-regularization inference: An elementary, general approach. Annual Reviews in Economics 7, 649–688.

Christensen, K., Siggaard, M., Veliyev, B., 2022. A machine learning approach to volatility forecasting. Journal of Financial Econometrics forthcoming.

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

Engle, R., 2002. New frontiers for ARCH models. Journal of Applied Econometrics 17, 425–446.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica 50, 987–1007.

Filipović, D., Khalilzadeh, A., 2021. Machine learning for predicting stock return volatility. Swiss Finance Institute Research Paper .

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Review of Financial Studies 33, 2326–2377.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2004. The MIDAS touch: Mixed data sampling regression models. Working Paper .

Ghysels, E., Santa-Clara, P., Valkanov, R., 2006. Predicting volatility: How to get most out of returns data sampled at different frequencies. Journal of Econometrics 131, 59–95.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Review of Financial Studies 33, 2223–2273.

Hansen, P.R., Huang, Z., Shek, H.H., 2012. Realized GARCH: A joint model for returns and realized measures of volatility. Journal of Applied Econometrics 27, 877–906.

Harris, L., 1986. A transaction data study of weekly and intradaily patterns in stock returns. Journal of Financial Economics 16, 99–117.

Jacod, J., 2008. Asymptotic properties of realized power variations and related functionals of semimartingales. Stochastic Processes and their Applications 118, 517–559.

Jacod, J., 2018. Limit of random measures associated with the increments of a Brownian semimartingale. Journal of Financial Econometrics 16.

Jacod, J., Li, Y., Mykland, P.A., Podolskij, M., Vetter, M., 2009. Microstructure noise in the continuous case: the pre-averaging approach. Stochastic Processes and their Applications 119, 2249–2276.

Jacod, J., Protter, P., 1998. Asymptotic error distributions for the euler method for stochastic differential equations. Annals of Probability , 267–307.

Judd, K.L., 1998. Numerical Methods in Economics. MIT Press.

Li, S.Z., Tang, Y., 2021. Forecasting realized volatility: An automatic system using many features and many machine learning algorithms. Available at SSRN 3776915 .

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

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

Mullainathan, S., Spiess, J., 2017. Machine learning: An applied econometric approach. Journal of Economic Perspectives 31, 87–106.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Patton, A., Zhang, H., 2022. Re-imgaing volatility: Computer vision approach for realized volatility forecasting. Working Paper .

Patton, A.J., Sheppard, K., 2009. Evaluating volatility and correlation forecasts, in: Handbook of Financial Time Series. Springer, pp. 801–838.

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

Patton, A.J., Weller, B.M., 2022. Risk price variation: The missing half of empirical asset pricing. Review of Financial Studies 35, 5127–5184.

Poon, S.H., Granger, C.W.J., 2003. Forecasting volatility in financial markets: A review. Journal of Economic Literature 41, 478–539.

Reisenhofer, R., Bayer, X., Hautsch, N., 2022. Harnet: A convolutional neural network for realized volatility forecasting. arXiv preprint arXiv:2205.07719 .

Robbins, H., Monro, S., 1951. A stochastic approximation method. Annals of Mathematical Statistics , 400–407.

Shephard, N., Sheppard, K., 2010. Realising the future: forecasting with high-frequency-based volatility (HEAVY) models. Journal of Applied Econometrics 25, 197–231.

Wood, R.A., McInish, T.H., Ord, J.K., 1985. An investigation of transactions data for nyse stocks. Journal of Finance 40, 723–739.

Zhang, L., Mykland, P.A., Aït-Sahalia, Y., 2005. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. Journal of the American Statistical Association 100, 1394–1411.