Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 25 (2009) 218–238
www.elsevier.com/locate/ijforecast

Optimal combinations of realised volatility estimators☆

Andrew J. Patton*, Kevin Sheppard

Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom
Oxford-Man Institute of Quantitative Finance, University of Oxford, Blue Boar Court, Oxford, OX1 4EH, United Kingdom

Abstract

Recent advances in financial econometrics have led to the development of new estimators of asset price variability using frequently-sampled price data, known as "realised volatility estimators" or simply "realised measures". These estimators rely on a variety of different assumptions and take many different functional forms. Motivated by the empirical success of combination forecasts, this paper presents a novel approach for combining individual realised measures to form new estimators of price variability. In an application to high frequency IBM price data over the period 1996–2008, we consider 32 different realised measures from 8 distinct classes of estimators. We find that a simple equally-weighted average of these estimators cannot generally be out-performed, in terms of accuracy, by any individual estimator. Moreover, we find that none of the individual estimators encompasses the information in all other estimators, providing further support for the use of combination realised measures.
© 2009 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

Keywords: Realised variance; Volatility forecasting; Forecast comparison; Forecast combination

1. Introduction

The development of new estimators of asset price variability has been an active area of econometric research in the past decade. These estimators, known as "realised volatility estimators" or "realised measures", exploit the information in high frequency data on asset prices (e.g., 5-min prices) to estimate the variability of the price process over a longer period, commonly one day. Older studies in this "realised volatility" literature, such as French, Schwert, and Stambaugh (1987), Merton (1980), and Zhou (1996), recognised the benefits from such an approach in increased accuracy, and recent work[1] has built on this to propose estimators that are more efficient, are robust to market microstructure effects, and can estimate the variation due to the continuous part of

☆ The web appendix for this paper is available at http://www.forecasters.org/ijf/.

* Corresponding author at: Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom.

E-mail addresses: andrew.patton@economics.ox.ac.uk (A.J. Patton), kevin.sheppard@economics.ox.ac.uk (K. Sheppard).

[1] See Andersen and Bollerslev (1998), Andersen, Bollerslev, Diebold, and Labys (2001a, 2003), Aït-Sahalia, Mykland, and Zhang (2005), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008, in press), Barndorff-Nielsen and Shephard (2002, 2004, 2006), Christensen and Podolskij (2007), Bandi and Russell (2006, 2008), Christensen, Oomen, and Podolskij (2008), Hansen and Lunde (2006a), Large (2005), Oomen (2006) and Zhang, Mykland, and Aït-Sahalia (2005) amongst others.

0169-2070/$ - see front matter © 2009 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2009.01.011

the price process separately from the variation due to the "jump" part of the price process. See Andersen, Bollerslev, Christoffersen, and Diebold (2006) and Barndorff-Nielsen and Shephard (2007) for recent reviews of this rapidly-evolving body of literature.

This paper seeks to answer the following simple question: do *combinations* of the above estimators offer gains in average accuracy relative to individual estimators? It has long been known in the forecasting literature that combinations of individual forecasts often out-perform even the best individual forecast, see Becker and Clements (2008), Bates and Granger (1969), Newbold and Granger (1974), and Stock and Watson (2004), for example, and see Clemen (1989) and Timmermann (2006) for reviews of this field.[2] Timmermann (2006) summarises three explanations for why combination forecasts work well in practice: they combine the information contained in each individual forecast; they average across differences in the way individual forecasts are affected by structural breaks; and they are less sensitive to possible mis-specification of individual forecasting models (see also Clements & Hendry, 1998, on forecast model mis-specification). Each of these three points applies equally to the problem of estimating price variability: individual realised measures use different pieces of information from high frequency data, they may be differently affected by structural breaks (caused by, for example, changes in the market microstructure), and they may be affected by mis-specification to various degrees. Thus, there is reason to believe that a combination realised measure may out-perform individual realised measures.

The theoretical contribution of this paper is to propose methods for constructing optimal combinations of realised measures, where optimality is formally defined below. The construction of combination estimators for asset price variability (measured by its quadratic variation, QV) differs in an important way from the usual forecast combination problem: the task is complicated by the fact that QV is not observed, even ex post. This means that measuring the accuracy of a given estimator of QV, or constructing a combination estimator that is as accurate as possible, has to be done using proxies (or, in our case, instruments) for the true latent QV. Our theoretical work extends the data-based method for estimating the relative accuracy of realised measures suggested by Patton (2008) to allow the estimation of optimal combination weights, or optimal combination functional forms more generally. Our methods use the time series aspect of the data (i.e., they are "large T"), which enables us to avoid making strong assumptions about the underlying price process, but at the cost of having to employ some assumptions (such as standard mixing and moment conditions) to ensure that a central limit theorem can be invoked.

The main contribution of this paper is to apply our combination methods to a collection of 32 different realised measures, across 8 distinct classes of estimators, estimated using high frequency data on IBM over the period 1996–2008. We present results on the ranking of the individual estimators and "simple" combination estimators such as the arithmetic mean, the geometric mean and the median, both over the full sample period and over three subsamples (1996–1999, 2000–2003, 2004–2008), using two distance measures, the mean squared error (MSE) and the QLIKE distance measure described below. We use the step-wise hypothesis testing method of Romano and Wolf (2005) to find the estimators that are significantly better (and significantly worse) than simple daily squared returns, the standard 5-min realised volatility estimator, or a simple equally-weighted average of all estimators. We also use the "model confidence set" (MCS) of Hansen, Lunde, and Nason (2005) to find the set of estimators that are not significantly different from the best estimator. Using the Romano–Wolf test, we find that only 2 of the 32 different realised measures significantly out-perform the simple average in the full sample, under QLIKE, but none significantly out-perform it under MSE. Many individual realised measures significantly under-perform the simple average estimator.

We also estimate optimal combination estimators, under both MSE and QLIKE, and examine which individual realised measures enter significantly into the optimal combination forecast, or enter with non-zero weight into an optimal constrained forecast. We find that weight is given to a variety of realised measures, including both simple and more sophisticated estimators. Importantly, we find that no

[2] See Halperin (1961) and Reid (1968) for interesting early work on combining different estimates of a mean, and different noisy estimates of GDP, as opposed to combining forecasts.

individual estimator encompasses the information in all other estimators, providing further support for the use of combination realised measures. Finally, we conduct an out-of-sample forecasting experiment to examine whether the gains in estimation accuracy carry over to gains in volatility forecast performance. Using the simple HAR model of Corsi (2004) to obtain one-step-ahead forecasts, we find that, unsurprisingly, better estimation accuracy generally leads to better forecast accuracy, although the rankings are not identical. We also find that a single forecast based on a combination *estimator* significantly out-performs a combination *forecast* based on many individual estimators.

2. Combining realised measures

2.1. Notation

The latent target variable, generally the quadratic variation (QV) or integrated variance (IV) of an asset price process, is denoted θ_t. We assume that θ_t is a $\mathcal{F}_t$-measurable scalar, where $\mathcal{F}_t$ is the information set generated by the complete path of the log-price process. The estimators ("realised measures" or "realised volatility estimators") of θ_t are denoted $X_{i,t}$, $i = 1, 2, \ldots, n$. In addition to being estimators with different functional forms, these may include the same type of estimator applied to data sampled at different frequencies (e.g., standard RV estimated on 1-min or 5-min data). Let $g(\mathbf{X}_t, \mathbf{w})$ denote a parametric combination estimator, where $\mathbf{w}$ is a finite-dimensional vector of parameters to be estimated from the data.

Defining an "optimal" combination estimator requires a measure of accuracy for a given estimator. Two popular measures in the volatility literature are the MSE and QLIKE measures:

$$\text{MSE} \quad L(\theta, X) = (\theta - X)^2 \tag{1}$$

$$\text{QLIKE} \quad L(\theta, X) = \frac{\theta}{X} - \log\left(\frac{\theta}{X}\right) - 1. \tag{2}$$

The QLIKE distance measure is a simple modification of the familiar Gaussian log-likelihood, with the modification being such that the minimum distance of zero is obtained when $X = \theta$. Our result below will be shown to hold for a more general class of distance measures, namely the class of "robust" pseudo-distance measures proposed by Patton (2006):

$$L(\theta, X) = \tilde{C}(X) - \tilde{C}(\theta) + C(X)(\theta - X), \tag{3}$$

with C being some function that is decreasing and twice-differentiable on the supports of both θ and X, and where $\tilde{C}$ is the anti derivative of C. In this class, each pseudo-distance measure L is completely determined by the choice of C, and MSE and QLIKE are obtained (up to location and scale constants) when $C(z) = -z$ and $C(z) = 1/z$ respectively. Finally, it is convenient to introduce the following quantities:

$$L^*(\mathbf{w}) \equiv E\left[L(\theta_t, g(\mathbf{X}_t, \mathbf{w}))\right] \tag{4}$$

$$\tilde{L}^*(\mathbf{w}) \equiv E\left[L(Y_t, g(\mathbf{X}_t, \mathbf{w}))\right] \tag{5}$$

$$\bar{L}_T(\mathbf{w}) \equiv \frac{1}{T}\sum_{t=1}^{T} L(Y_t, g(\mathbf{X}_t, \mathbf{w})), \tag{6}$$

where the dependence of $\bar{L}_T, \tilde{L}^*$ and L^* on the function g is suppressed for simplicity. Y_t is an observable proxy for the latent θ_t, and is further discussed in the following section.

2.2. Estimating optimal combinations of realised measures

In this section we provide the theory underlying the estimation of optimal combination estimators, building on the work of Patton (2008), who considered rankings of realised measures. The ranking method of Patton (2008) provides a means of consistently estimating the difference in average accuracy of two competing estimators. This method is based on an instrumental variables-type approach, which overcomes both the latent nature of the target variable (θ_t), and problems arising from correlations between the errors in the competing estimators (X_{it}) and the proxy $\left(\tilde{\theta}_t\right)$ for the latent target variable. We will denote a generic combination estimator as $g(\mathbf{X}_t, \mathbf{w})$. A concrete example of a combination estimator is the linear combination:

$$g_L(\mathbf{X}_t, \mathbf{w}) = \omega_0 + \sum_{i=1}^{n} \omega_i X_{it}. \tag{7}$$

In volatility applications, multiplicative forecast combinations may also be used:

$$g_M(\mathbf{X}_t, \mathbf{w}) = \omega_0 \times \prod_{i=1}^{n} X_{it}^{\omega_i}. \tag{8}$$

We will define the optimal combination parameter, $\mathbf{w}^*$, the feasible optimal combination parameter, $\tilde{\mathbf{w}}^*$, and the estimated combination parameter, $\hat{\mathbf{w}}^*_T$, as follows:

$$\mathbf{w}^* \equiv \arg\min_{\mathbf{w}\in\mathcal{W}} L^*(\mathbf{w}),$$
$$\tilde{\mathbf{w}}^* \equiv \arg\min_{\mathbf{w}\in\mathcal{W}} \tilde{L}^*(\mathbf{w}),$$
$$\hat{\mathbf{w}}^*_T \equiv \arg\min_{\mathbf{w}\in\mathcal{W}} \bar{L}_T(\mathbf{w}), \quad (9)$$

where $\mathcal{W}$ is the parameter space and L^*, $\tilde{L}^*$, and $\bar{L}_T$ are defined in Eqs. (4) to (6).

The difficulty in estimating optimal combinations of realised measures lies in the fact that quadratic variation is not observable, even ex post. Thus, measuring the accuracy of a given estimator, or constructing a combination estimator that is as accurate as possible, is not straightforward. This is related to the problem of volatility forecast evaluation and comparison, where the target variable is also unobservable. Andersen and Bollerslev (1998), Andersen, Bollerslev, and Meddahi (2005), Hansen and Lunde (2006b), Meddahi (2001), Patton (2006), and Patton and Sheppard (in press) discuss this problem in the context of volatility forecasting. Unfortunately, the methods developed for volatility forecasting are not directly applicable to the problem of evaluating realised measure accuracy, or the construction of combinations of realised measures, due to a difference in the information set that is used: in volatility forecasting applications, the estimate of θ_t will be based on $\mathcal{F}_{t-1}$, while in volatility *estimation* applications, the estimate of θ_t will be based on $\mathcal{F}_t$. As discussed by Patton (2008), this subtle change in information sets forces a substantial change in methods for comparing and combining realised measures. Ignoring this change leads to combination estimators that are biased and inconsistent. These problems arise because the error in the proxy for θ_t is correlated with the error in the estimator, a problem that does not arise in volatility forecasting applications, under basic assumptions.

Following Patton (2008), we will consider the case where an unbiased proxy for θ_t is known to be available. An example of this is the daily squared returns, which can plausibly be assumed to be free from microstructure and other biases, and so is an unbiased, albeit noisy, estimator of QV.

Assumption P1. $\tilde{\theta}_t = \theta_t + \nu_t$, with $E\left[\nu_t|\mathcal{F}_{t-1},\theta_t\right] = 0$.

Next, we need an assumption about the dynamics of the target variable θ_t. Patton (2008) considers two assumptions here, either that θ_t follows a (possibly heteroskedastic) random walk, or that θ_t follows a stationary AR(p) process.[3] For the high frequency IBM data studied below, Patton (2008) found that the random walk approximation was satisfactory, and so for simplicity we will focus on that case; the extension to the AR(p) approximation is straightforward.

Assumption T1. $\theta_t = \theta_{t-1} + \eta_t$, with $E\left[\eta_t|\mathcal{F}_{t-1}\right] = 0$.

In order to overcome the problem of correlated measurement errors in $\tilde{\theta}_t$ and X_{it}, Patton (2008) suggests using a lead, or combination of leads, of $\tilde{\theta}_t$ in the estimation of the optimal combination weights. Denoting this as Y_t, we make the following assumption:

Assumption P2. $Y_t = \sum_{j=1}^{J}\lambda_j\tilde{\theta}_{t+j}$, where $1 \leq J < \infty$, $\lambda_j \geq 0\ \forall\, j$ and $\sum_{j=1}^{J}\lambda_j = 1$.

In practice, there is a trade-off to be made in choosing J and λ_j. If the random walk Assumption (T1) was literally true, then the optimal choice would be to make the value of J large and use exponentially-declining weights, see Muth (1960). If the random walk assumption is merely an approximation, then using fewer lags is likely to lead to a better approximation than using longer lags, at the cost of a noisier instrument. A simple and conservative choice for Y_t, and the one we adopt in our empirical work below, is to set $Y_t = \tilde{\theta}_{t+1}$.

To obtain the asymptotic distribution of the estimated optimal combination weights, we require assumptions sufficient for a central limit theorem to hold. Several different sets of assumptions may be employed here; we use the high-level assumptions of Gallant and White (1988), and refer the interested reader there for more primitive assumptions that

[3] The need for an assumption about the dynamics of the target variable comes from the use of a lead of the proxy, $\tilde{\theta}_t$, see Assumption P2, and the non-linear nature of the quantity being estimated, namely the difference in average distance to the target variable.

may be used for non-linear, dynamic m-estimation problems such as ours. See also Davidson and MacKinnon (1993) for a concise and very readable overview of asymptotic normality for m-estimators.

Assumption A1(a). $\bar{L}_T(\mathbf{w}) - \tilde{L}^*(\mathbf{w}) \xrightarrow{p} 0$ uniformly on $\mathcal{W}$.

Assumption A1(b). $\tilde{L}^*(\mathbf{w})$ has a unique minimiser $\tilde{\mathbf{w}}^*$.

Assumption A1(c). g is twice continuously differentiable with respect to $\mathbf{w}$.

Assumption A1(d). Let $A_T(\mathbf{w}) \equiv \nabla_{\mathbf{ww}} \bar{L}_T(\mathbf{w})$, then $A_T(\mathbf{w}) - A(\mathbf{w}) \xrightarrow{p} 0$ uniformly on $\mathcal{W}$, where $A(\mathbf{w})$ is a finite positive definite matrix of constants for all $\mathbf{w} \in \mathcal{W}$.

Assumption A1(e). $T^{-1/2} \sum_{t=1}^{T} \nabla_{\mathbf{w}} L(Y_t, g(\mathbf{X}_t, \mathbf{w})) \xrightarrow{D} N(0, B(\mathbf{w}))$, where $B(\mathbf{w})$ is finite and positive definite for all $\boldsymbol{w} \in \mathcal{W}$.

With the above assumptions in hand, we now present our main theoretical result.

Proposition 1. *If the distance measure L is a member of the class in Eq. (3), and if $\tilde{\mathbf{w}}^*$ is interior to $\mathcal{W}$, then under Assumptions P1, P2, T1 and A1, we have:*

$$\hat{V}_T^{-1/2} \sqrt{T} \left(\hat{\mathbf{w}}_T^* - \mathbf{w}^*\right) \xrightarrow{D} N(0, I)$$

where $\hat{V}_T \equiv \hat{A}_T^{-1} \hat{B}_T \hat{A}_T^{-1}$,

$$\hat{A}_T \equiv \frac{1}{T} \sum_{t=1}^{T} \nabla_{\mathbf{ww}} L\left(Y_t, g\left(\mathbf{X}_t, \hat{\mathbf{w}}_T^*\right)\right),$$

$$B_T \equiv V\left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla_{\mathbf{w}} L\left(Y_t, g\left(\mathbf{X}_t, \hat{\mathbf{w}}_T^*\right)\right)\right]$$

and $\hat{B}_T$ is some symmetric and positive definite estimator of B_T such that $\hat{B}_T - B_T \xrightarrow{p} 0$.

Proof. We first show that $\mathbf{w}^* = \tilde{\mathbf{w}}^*$. This part of the proof is a corollary to Proposition 2(a) of Patton (2008). Consider a second-order mean-value expansion of $L(Y_t, g(\mathbf{X}_t, \mathbf{w}))$ around θ_t:

$$L(Y_t, g(\mathbf{X}_t, \mathbf{w})) = L(\theta_t, g(\mathbf{X}_t, \mathbf{w})) + \frac{\partial L(\theta_t, g(\mathbf{X}_t, \mathbf{w}))}{\partial \theta}(Y_t - \theta_t) + \frac{1}{2} \frac{\partial^2 L\left(\ddot{\theta}_t, g(\mathbf{X}_t, \mathbf{w})\right)}{\partial \theta^2}(Y_t - \theta_t)^2$$

where $\ddot{\theta}_t = \delta_t Y_t + (1 - \delta_t)\theta_t$ for some $\delta_t \in [0, 1]$. Under Assumptions P1, P2 and T1, Patton (2008) shows that the second term in this expansion has mean zero, and so we obtain:

$$E[L(Y_t, g(\mathbf{X}_t, \mathbf{w}))] = E[L(\theta_t, g(\mathbf{X}_t, \mathbf{w}))] + \frac{1}{2} E\left[\frac{\partial^2 L\left(\ddot{\theta}_t, g(\mathbf{X}_t, \mathbf{w})\right)}{\partial \theta^2}(Y_t - \theta_t)^2\right].$$

Distance measures in the class in Eq. (3) yield $\partial^2 L(\theta, X) / \partial \theta^2 = -C'(\theta)$, and so

$$E[L(Y_t, g(\mathbf{X}_t, \mathbf{w}))] = E[L(\theta_t, g(\mathbf{X}_t, \mathbf{w}))] - \frac{1}{2} E\left[C'\left(\ddot{\theta}_t\right)(Y_t - \theta_t)^2\right].$$

Notice that the second term above does not depend on $\mathbf{w}$, and thus the parameter that minimises $E[L(Y_t, g(\mathbf{X}_t, \mathbf{w}))]$ is the same as that which minimises $E[L(\theta_t, g(\mathbf{X}_t, \mathbf{w}))]$. Thus $\tilde{\mathbf{w}}^* = \mathbf{w}^*$.

Next we obtain the asymptotic distribution of $\hat{\mathbf{w}}_T^*$. This part of the proof uses standard results from m-estimation theory; see Gallant and White (1988), for example. Under Assumption A1(a) and A1(b), Theorem 3.3 of Gallant and White (1988) yields $\hat{\mathbf{w}}_T^* - \tilde{\mathbf{w}}^* \xrightarrow{p} 0$. Combining this with the fact that $\tilde{\mathbf{w}}^* = \mathbf{w}^*$ yields consistency of $\hat{\mathbf{w}}_T^*$ for the parameter of interest: $\hat{\mathbf{w}}_T^* - \mathbf{w}^* \xrightarrow{p} 0$. Under Assumption A1, Theorem 5.1 of Gallant and White (1988) yields the asymptotic normality of $\hat{\mathbf{w}}_T^*$, centered around $\tilde{\mathbf{w}}^*$. Combining this with $\tilde{\mathbf{w}}^* = \mathbf{w}^*$ from above yields the desired result. ■

The above proposition shows that it is possible to consistently estimate the optimal combination weights from the data, by employing a "robust" loss function of the form in Eq. (3), and using a lead (or a combination of leads) of a conditionally unbiased proxy for θ_t. This proposition further shows how to compute standard errors on these estimated optimal weights. The use of a proxy, Y_t, for the true quadratic variation, θ_t, means that these standard errors will generally be larger than those that would be obtained if θ_t was observable; nevertheless, these standard errors can be estimated using the expressions above.

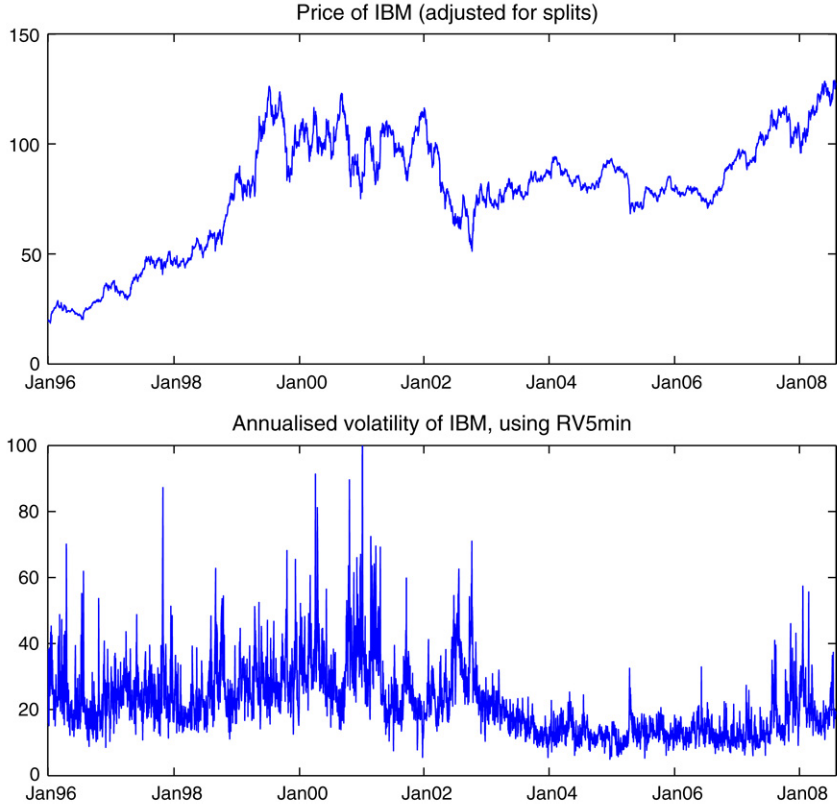

Fig. 1. This figure plots IBM price and volatility over the period January 1996 to July 2008. The price is adjusted for stock splits, and the volatility is computed using realised volatility based on 5-min calendar-time trade prices, annualised using the formula $\sigma_t = \sqrt{252 \times RV_t}$.

3. Application to estimating stock return volatility

3.1. Data description

In this section we consider the problem of estimating the quadratic variation of the open-to-close (9:45am to 4pm) continuously-compounded return on IBM, using a variety of different estimators and sampling frequencies. We use data on NYSE trade prices from the TAQ database over the period from January 1996 to July 2008, yielding a total of 3168 daily observations.[4] Fig. 1 reveals that the sample includes periods of rising prices and moderate-to-high volatility (roughly 1996–1999), of slightly falling prices and relatively high volatility (roughly 2000–2003), and of mostly stable prices and relatively low volatility (2004–2008). In addition to considering the full sample estimates of optimal combination estimators, we will consider the results for each of these three sub-samples.

3.2. Description of the individual estimators

The motivation for our study of realised measures is that the various forms of realised measures that have been proposed in the literature to date, and the different pieces of information captured by each, may

[4] We use trade prices from the NYSE only, between 9:45am and 4:00pm, with a *g127* code of 0 or 40, a *corr* code of 0 or 1, positive *size*, and *cond* **not** equal to "O", "Z", "B", "T", "L", "G", "W", "J", or "K". Further, the data were cleaned for outliers and related problems (e.g. prices of zero were dropped). The average proportion of observations lost each day by such cleaning was 0.28%, i.e., just over one quarter of one percent. Further, if more than one price was observed with the same time stamp then we used the median of these prices. See Barndorff-Nielsen et al. (in press) for a discussion of cleaning high frequency data.

allow for the construction of combination estimators that out-perform any given individual estimator. With this in mind, we consider a large collection of different realised measures. We follow the implementation of the authors of the original paper as closely as possible (and in most cases, exactly). We omit detailed definitions and descriptions of each estimator in the interests of space, and instead refer the interested reader to the original papers.

We firstly consider the standard realised variance, defined as:

$$RV_t^{(m)} = \sum_{j=1}^{m} r_{t,j}^2 \tag{10}$$

where m is the number of intra-daily returns used, and $r_{t,j}$ is the jth return on day t. The number of intra-daily returns used can vary between 22,500 (if we sample prices every second between 9:45am and 4pm) and 1 (if we sample just the open and close prices). To keep the number of estimators tractable, and with the high degree of correlation between RV estimators with similar sampling frequencies in mind, we select six sampling frequencies: 1 s, 5 s, 1 min, 5 min, 62.5 min (which we will abbreviate as "1 h"), and 1 trade day (375 min). The first set of RV estimators is based on prices sampled in "calendar time" using last-price interpolation, meaning that we construct a grid of times between 9:45am and 4pm with the specified number of minutes between each point, and use the most recent price as the one for a given grid point.

Next we consider RV estimators computed using the same formula, but with prices sampled in "tick time" (also known as "business time" or "trade time"). In this sampling scheme, a price series for each day is constructed by skipping every x trades: this leads to prices that are evenly spaced in "event time", but generally not in calendar time. If the trade arrival rate is correlated with the level of volatility, then tick-time sampling produces high-frequency returns which are approximately homoskedastic. Theory suggests that this should improve the accuracy of RV estimation, see Hansen and Lunde (2006a) and Oomen (2006). We consider average sampling frequencies in tick-time that correspond to those used in calendar time: 1 s, 5 s, 1 min, 5 min, 62.5 min and 375 min. The highest and lowest of these frequencies lead to estimators that are numerically identical to calendar-time RV, and so we drop these from the analysis.

We then draw on the work of Bandi and Russell (2006, 2008), who provide a method of estimating the optimal (calendar-time) sampling frequency, for each day, for realised variance in the presence of market microstructure noise. This formula relies on estimates of the variance and kurtosis of the microstructure noise, as well as preliminary estimates of the integrated variance (IV) and integrated quarticity (IQ) of the price process. We follow Bandi and Russell (2008), who also study IBM stock returns, and estimate the moments of the microstructure noise using 1s returns, and use 15-min returns to obtain preliminary estimates of the IV and IQ. Bandi and Russell (2008) also propose a bias-corrected realised variance estimator, which removes the estimated impact of the microstructure noise; we consider the Bandi-Russell RV estimator both with $\left(\mathrm{RV}^{BR,bc}\right)$ and without $\left(\mathrm{RV}^{BR}\right)$ this bias correction.

Our second class of realised volatility estimators is the first-order autocorrelation-adjusted RV estimator (RV^{AC1}) presented by French et al. (1987) and Zhou (1996), and studied by Bandi and Russell (2008) and Hansen and Lunde (2006a), amongst others. We implement this estimator on 1-min and 5-min prices sampled in calendar time.

Our third class of estimators includes the two-scale estimator (TSRV) of Zhang et al. (2005) and the multi-scale estimator (MSRV) of Zhang (2006). As their names suggest, these estimators use realised variances computed using more than one sampling frequency, which is shown, under certain conditions, to lead to consistency of the estimator in the presence of noise, and to efficiency gains. Following the theoretical suggestions in those papers, we implement these estimators at the highest possible frequency (1 tick), and for comparison we also implement them on one-minute tick-time prices.

The fourth set of estimators are the "realised kernels" (RK) of Barndorff-Nielsen et al. (2008), BNHLS henceforth. This is a broad class of estimators, which nests the RV^{AC1} estimator, and we consider several variations. Firstly, we consider RK with the Bartlett kernel, as this estimator was shown by BNHLS to be asymptotically equivalent to TSRV. Second, we consider RK with the "cubic" kernel, which was shown to be asymptotically equivalent to MSRV. For both RK^{bart} and RK^{cubic} we consider both 1-tick sampling and 1-min tick-time sampling, and in all cases we use the optimal bandwidth for a given kernel, as provided by Barndorff-Nielsen et al. (2008).[5] Next we consider the "modified Tukey-Hanning$_2$" (TH2) kernel, following their empirical application to General Electric stock returns. They suggest using 1-min tick-time sampling, which we implement here, and we also implement 1-tick sampling for comparison. Finally, we consider the "non-flat-top Parzen" kernel of Barndorff-Nielsen et al. (in press), which is designed to guarantee non-negativity of the estimator (which is not ensured for the other RK estimators considered above). We implement this with their optimal bandwidth formula, and, following their application to Alcoa stock returns, use 1-min tick-time sampling, as well as 1-tick sampling for comparison.

Our fifth type of realised measure is the "realised range-based variance" (RRV) of Christensen and Podolskij (2007) and Martens and van Dijk (2007). We follow Christensen and Podolskij's implementation of this estimator and use 5-min blocks. Rather than estimate the number of prices to use within each block from the number of non-zero return changes, as was done by Christensen and Podolskij (2007), we simply use 1-min prices within each block, giving us 5 prices per block, compared with around 7 in their application to General Motors stock returns. We implement RRV using tick-time sampling.

The next three types of estimators we consider estimate the part of the quadratic variation that is due to the continuous semimartingale, that is, the integrated variance, IV. The previous five types of estimators all estimate QV, which differs from IV in the presence of jumps. If jumps are unpredictable (or less predictable than IV), as was found by Andersen, Bollerslev, and Diebold (2007), then there may be benefits in using estimators that focus on IV rather than QV in forecasting applications.

The sixth set of realised measures is the "bi-power variation" (BPV) of Barndorff-Nielsen and Shephard (2006). These authors implemented their estimator using 5-min calendar-time returns; however, this was presumably partially dictated by their data (indicative quotes for the US dollar/German Deutsche mark and US dollar/Japanese yen exchange rates), for which this was the highest frequency. We thus implement BPV at both the 5-min and 1-min frequencies, using calendar-time sampling.

Our seventh class of realised measures is the "quantile-based realised variance" (QRV) of Christensen et al. (2008). For implementation, we follow Christensen et al.'s application to Apple stock returns, and use quantiles of 0.85, 0.90 and 0.96, and prices sampled every 1 min, using tick-time sampling, with the number of subintervals ("n" in their notation) set to one.

Our eighth and final class of realised measures is the MedRV and MinRV estimators of Andersen, Dobrev, and Schaumburg (2008), which were designed to overcome some of the practical difficulties suffered by BPV, and also to provide some robustness to market microstructure noise. Following the empirical results presented by Andersen et al. (2008), we implement these estimators using 1-min tick-time sampling.

In total, we have 32 different realised measures, from 8 different classes of estimators, with a variety of sampling frequencies and sampling schemes. To the best of our knowledge, this is the largest collection of realised measures considered in a single empirical study to date.

Table 1 presents some summary statistics for these 32 estimators. $RV^{BR,bc}$ has the smallest average value, 1.425, and $RV^{1\,s}$ has the largest average, 3.158. More familiar estimators, such as $RV^{1\,day}$ and $RV^{5\,min}$, have average values of around 2.4, corresponding to 24.6% annualised standard deviation. Whilst RV^{1day} has a reasonable average value, it performs poorly on the other summary statistics: it has the highest standard deviation, skewness, and kurtosis of all 32 estimators. $RV^{BR,bc}$ is the estimator with the lowest standard deviation, while QRV has the lowest skewness and kurtosis. RV^{1day} and $RV^{BR,bc}$ are the only estimators which generate estimates of QV that are not strictly positive: RV^{1day}'s minimum value is zero, which it attains on 32 days (around 1% of the sample), while $RV^{BR,bc}$'s minimum value is −8.10, and it is non-positive on 68 days (around 2.1% of the sample). The bias-correction term in this estimator is clearly too large on these days, causing the estimator to go below zero. None of the other estimators have a minimum value that is non-positive (including the

[5] These optimal bandwidths, like the RV^{BR} sampling frequencies, are estimated for each day in the sample, and so can change with market conditions, the level of market microstructure noise, etc.

Table 1
Summary statistics of the realised measures.

	Standard mean	Deviation	Skewness	Kurtosis	Minimum
$RV^{1\,s}$	3.158	3.005	2.940	22.270	0.168
$RV^{5\,s}$	3.023	2.853	2.998	23.977	0.157
$RV^{1\,min}$	2.438	2.387	3.647	34.193	0.116
$RV^{5\,min}$	2.366	2.901	6.863	113.221	0.098
RV^{1h}	2.307	3.699	7.315	111.988	0.018
RV^{1day}	2.403	6.228	10.638	193.816	0.000
$RV^{tick,5\,s}$	3.147	3.010	2.940	22.220	0.169
$RV^{tick,1\,min}$	2.427	2.425	4.019	42.059	0.107
$RV^{tick,5\,min}$	2.383	2.912	7.060	118.930	0.072
$RV^{tick,1h}$	2.346	3.813	7.588	110.325	0.014
RV^{BR}	2.345	2.511	3.711	31.423	0.079
$RV^{BR,bc}$	1.425	1.573	5.942	100.202	−8.103
$RV^{AC1,1\,min}$	2.440	2.392	3.592	32.743	0.117
$RV^{AC1,5\,min}$	2.363	2.896	6.702	107.095	0.098
$TSRV^{tick}$	2.177	2.202	3.994	39.384	0.081
$TSRV^{tick,1\,min}$	2.398	2.947	7.501	141.857	0.084
$MSRV^{tick}$	2.181	2.287	5.572	85.553	0.081
$MSRV^{tick,1\,min}$	2.441	2.972	7.771	152.748	0.112
RK^{bart}	2.362	2.747	7.618	153.602	0.104
$RK^{bart,1\,min}$	2.362	2.963	7.103	121.566	0.073
RK^{cubic}	2.409	2.910	7.473	143.174	0.119
$RK^{cubic,1\,min}$	2.258	2.950	6.751	100.064	0.061
RK^{TH2}	2.381	2.784	7.761	158.827	0.109
$RK^{TH2,1\,min}$	2.332	2.970	7.000	113.749	0.062
RK^{NFP}	2.361	2.932	7.214	126.027	0.094
$RK^{NFP,1\,min}$	2.257	2.931	6.424	87.273	0.052
RRV	2.310	2.537	4.647	52.061	0.123
$BPV^{1\,min}$	2.105	2.075	2.632	12.867	0.077
$BPV^{5\,min}$	2.202	2.563	3.945	29.413	0.099
QRV	2.441	2.273	2.430	11.563	0.104
MedRV	2.260	2.157	2.600	13.216	0.109
MinRV	2.226	2.156	2.583	12.697	0.120

Notes: This table presents basic summary statistics on the 32 different realised measures considered in this paper.

TSRV, MSRV and RK estimators, which do not ensure non-negativity of the estimates).[6]

Table 2 presents a subset of the correlation matrix of these estimators. We present the correlation of each estimator with two standard estimators in the literature ($RV^{5\,min}$ and RV^{1day}), a naïve choice given high frequency data ($RV^{1\,s}$), an early modification of the standard RV ($RV^{AC1,1\,min}$), and two recently-proposed estimators ($RK^{TH2,1\,min}$ and QRV). This table shows that these estimators are generally highly correlated, which is to be expected, since they are all influenced by the long-run component of IBM volatility. The average correlation across all elements of their correlation matrix is 0.879. This should be kept in mind when interpreting the estimated optimal combination weights in Section 3.4. The highest correlation between any two estimators is between $RV^{1\,s}$ and $RV^{tick,5\,s}$, which is 0.9998. The lowest correlation between any two estimators is between RV^{1day} and $RV^{BR,bc}$, at 0.314.

[6] Before ranking and averaging the estimators in the following section, we put them through a simple "insanity filter": if an estimator had a value less than 0.001 on a given day, that value was replaced with the value of the estimator on the previous day. As Table 1 reveals, this insanity filter was only needed for RV^{1day} and $RV^{BR,bc}$. This filter is required for the use of the QLIKE distance measure, which assumes that the estimators are all strictly positive.

Table 2
Correlation between the realised measures.

	$RV^{1\,s}$	$RV^{5\,min}$	RV^{1day}	$RV^{AC1,1\,min}$	$RK^{TH2,1\,min}$	QRV
$RV^{1\,s}$	1	0.855	0.431	0.939	0.839	0.906
$RV^{5\,s}$	0.999	0.869	0.441	0.950	0.853	0.915
$RV^{1\,min}$	0.938	0.948	0.517	0.997	0.939	0.956
$RV^{5\,min}$	0.855	1	0.570	0.947	0.985	0.872
RV^{1h}	0.643	0.788	0.728	0.743	0.804	0.712
RV^{1day}	0.431	0.570	1	0.514	0.593	0.483
$RV^{tick,5\,s}$	1.000	0.855	0.431	0.938	0.838	0.905
$RV^{tick,1\,min}$	0.935	0.951	0.522	0.993	0.943	0.947
$RV^{tick,5\,min}$	0.852	0.979	0.583	0.944	0.988	0.875
$RV^{tick,1h}$	0.653	0.802	0.697	0.751	0.815	0.711
RV^{BR}	0.902	0.951	0.522	0.978	0.951	0.936
$RV^{BR,bc}$	0.794	0.734	0.314	0.820	0.711	0.744
$RV^{AC1,1\,min}$	0.939	0.947	0.514	1	0.939	0.955
$RV^{AC1,5\,min}$	0.854	0.997	0.576	0.948	0.986	0.874
$TSRV^{tick}$	0.913	0.938	0.507	0.983	0.931	0.935
$TSRV^{tick,1\,min}$	0.866	0.979	0.559	0.958	0.982	0.881
$MSRV^{tick}$	0.904	0.952	0.519	0.980	0.945	0.913
$MSRV^{tick,1\,min}$	0.859	0.979	0.570	0.957	0.984	0.877
RK^{bart}	0.877	0.974	0.546	0.968	0.973	0.888
$RK^{bart,1\,min}$	0.849	0.986	0.580	0.947	0.998	0.874
RK^{cubic}	0.862	0.980	0.564	0.958	0.985	0.879
$RK^{cubic,1\,min}$	0.818	0.976	0.604	0.921	0.993	0.854
RK^{TH2}	0.874	0.975	0.550	0.967	0.974	0.885
$RK^{TH2,1\,min}$	0.839	0.985	0.593	0.939	1	0.867
RK^{NFP}	0.854	0.986	0.582	0.951	0.997	0.876
$RK^{NFP,1\,min}$	0.819	0.973	0.613	0.921	0.990	0.859
RRV	0.902	0.984	0.550	0.982	0.974	0.933
$BPV^{1\,min}$	0.912	0.878	0.478	0.960	0.871	0.974
$BPV^{5\,min}$	0.848	0.952	0.533	0.932	0.935	0.914
QRV	0.906	0.872	0.483	0.955	0.867	1
MedRV	0.919	0.882	0.487	0.961	0.874	0.980
MinRV	0.917	0.873	0.478	0.954	0.864	0.973

Notes: This table presents a sub-set of the correlation matrix of the 32 different realised measures considered in this paper. The estimators in the columns correspond to standard choices in the extant literature (RV^{1day} and $RV^{5\,min}$), a naïve choice given high frequency data ($RV^{1\,s}$), and three other estimators from our empirical analysis ($RV^{AC1,1\,min}$, $RK^{TH2,1\,min}$ and QRV).

3.3. Results using simple combination estimators

In Table 3 we present the first set of empirical results of the paper. These tables present the estimated accuracy of each of the estimators using the ranking methodology of Patton (2008). We use the QLIKE distance measure for the analysis below, and report corresponding results using the MSE distance measure in a web appendix to this paper.[7] We use the random walk approximation (Assumption T1), with a one-period lead of the $RV^{5\,min}$ as the instrument for the latent quadratic variation to obtain these estimates.

The ranking method of Patton (2008) can only estimate the accuracy of an estimator relative to some other estimator, and in Table 3 we use $RV^{5\,min}$ as the base estimator; this choice is purely a normalisation and has no effect on the conclusions. Negative values in the first columns of Table 3 indicate that a given

[7] The main conclusions of this paper hold under both the MSE and QLIKE distance measures. The results under MSE are less precise, however, due to the heteroskedastic nature of volatility estimation. This leads to lower power in tests using this distance measure, see Patton (2006) and Patton and Sheppard (in press), for example.

Table 3
Performance of the realised measures.

	Avg ΔQLIKE	Rank				In MCS?			
	Full	Full	96–99	00–03	04–08	Full	96–99	00–03	04–08
$RV^{1\,s}$	−0.013	14	6	28	21	–	–	–	–
$RV^{5\,s}$	−0.021	7	5	25	11	–	✓	–	–
$RV^{1\,min}$	−0.040	2	3	3	1	✓	✓	✓	✓
$RV^{5\,min}$	0	23	13	23	25	–	–	–	–
RV^{1h}	0.596	33	33	34	33	–	–	–	–
RV^{1day}	29.191	35	35	35	35	–	–	–	–
$RV^{tick,5\,s}$	−0.014	13	7	29	17	–	–	–	–
$RV^{tick,1\,min}$	−0.039	3	1	1	5	✓	✓	✓	–
$RV^{tick,5\,min}$	−0.010	16	10	21	24	–	–	–	–
$RV^{tick,1h}$	0.526	32	32	33	34	–	–	–	–
RV^{BR}	−0.004	18	21	19	14	–	–	–	–
$RV^{BR,bc}$	1.126	34	34	32	32	–	–	–	–
$RV^{AC1,1\,min}$	−0.040	1	2	2	2	✓	✓	✓	✓
$RV^{AC1,5\,min}$	0.001	25	14	22	27	–	–	–	–
$TSRV^{tick}$	−0.001	22	27	15	20	–	–	–	–
$TSRV^{tick,1\,min}$	−0.004	19	20	16	22	–	–	–	–
$MSRV^{tick}$	−0.003	21	26	17	19	–	–	–	–
$MSRV^{tick,1\,min}$	−0.005	17	25	14	12	–	–	–	–
RK^{bart}	−0.015	11	17	10	7	–	–	–	–
$RK^{bart,1\,min}$	0.006	26	18	24	26	–	–	–	–
RK^{cubic}	−0.004	20	24	18	13	–	–	–	–
$RK^{cubic,1\,min}$	0.051	31	29	30	31	–	–	–	–
RK^{TH2}	−0.014	12	16	11	9	–	–	–	–
$RK^{TH2,1\,min}$	0.015	27	22	26	28	–	–	–	–
RK^{NFP}	0.001	24	23	20	23	–	–	–	–
$RK^{NFP,1\,min}$	0.044	30	28	31	30	–	–	–	–
RRV	−0.016	9	15	8	16	–	–	–	–
$BPV^{1\,min}$	0.029	28	31	12	8	–	–	–	–
$BPV^{5\,min}$	0.040	29	30	27	29	–	–	–	–
QRV	−0.035	4	4	5	4	–	✓	✓	–
MedRV	−0.024	6	9	6	10	–	–	–	–
MinRV	−0.010	15	19	9	15	–	–	–	–
RV^{Mean}	−0.030	5	8	4	3	–	–	✓	–
$RV^{Geo\text{-}mean}$	−0.015	10	12	13	18	–	–	–	–
RV^{Median}	−0.020	8	11	7	6	–	–	–	–

Notes: The first column of this table presents the average difference in QLIKE distance of each realised measure, relative to $RV^{5\,min}$, with negative (positive) values indicating that the estimator was on average closer to (further from) the target variable than $RV^{5\,min}$. Columns 2–5 present the rank of each estimator using the QLIKE distance, for the full sample period and for three sub-samples, 1996–1999, 2000–2003 and 2004–2008. The most accurate estimator is ranked 1, and the least accurate estimator is ranked 35. Columns 6–9 present an indicator of whether the estimator was in the "model confidence set" at the 90% confidence level (equal to ✓ if in, – if not) in the full sample and each of the three sub-samples.

estimator has a lower average distance to the latent QV (i.e., greater accuracy) than $RV^{5\,min}$, while positive values indicate a higher average distance than $RV^{5\,min}$.

We consider the 32 individual realised measures discussed in the previous section, as well as three simple combination estimators: the equally-weighted arithmetic mean, the equally-weighted geometric mean, and the median, leading to a total of 35 estimators. The most accurate estimator of QV is the simple $RV^{AC1,1\,min}$, which is ranked in the top 2 in all three sub-periods. The top 5 estimators in the full sample are $RV^{AC1,1\,min}$ (top), $RV^{1\,min}$ (second),

$\text{RV}^{tick,1\,\text{min}}$ (third), QRV (fourth) and RV^{Mean} (fifth). It is interesting that two of the top five estimators are simple RV applied to one-minute returns (in tick-time and calendar-time). Also noteworthy is the fact that simple combination estimators perform well: under QLIKE, the simple mean and median estimators are both in the top ten, with the mean being ranked fifth on the full sample. (Under the MSE the simple combination estimators are all ranked around 10th, with the best being the geometric mean, see Table 3A in the web appendix.)

The discussion of rankings of average accuracy is a useful initial look at the results, but a more formal analysis is desirable. We use two approaches. The first is the "model confidence set" (MCS) of Hansen et al. (2005), which was developed to obtain a set of forecasting models that contains the true "best" model out of the entire set of forecasting models with some specified level of confidence. It allows the researcher to identify the sub-set of models that are "not significantly different" from the unknown true best model. Patton (2008) shows that this methodology may be adapted to the problem of identifying the most accurate realised measures, under the assumptions discussed in Section 2. The last four columns of Table 3 show the results of the MCS procedure on the full sample and on three sub-samples.[8] Under QLIKE, the full-sample MCS, at the 90% confidence level, contains just 3 estimators: $\text{RV}^{1\,\text{min}}$, $\text{RV}^{tick,1\,\text{min}}$ and $\text{RV}^{AC1,1\,\text{min}}$, and does not include either more sophisticated estimators or the simple combination estimators.[9, 10]

The second formal analysis of the individual estimators and simple combination estimators uses the stepwise multiple testing method of Romano and Wolf (2005). This method identifies the estimators that are significantly either better or worse than a given benchmark estimator, while controlling the family-wise error rate of the complete set of hypothesis tests.[11] We consider three choices of benchmark estimator: $\text{RV}^{1\text{day}}$, which is the standard estimator in the absence of high frequency data; $\text{RV}^{5\,\text{min}}$, which is based on a rule-of-thumb from earlier papers in the RV literature (see Andersen, Bollerslev, Diebold, & Ebens, 2001b, and Barndorff-Nielsen & Shephard, 2002, for example); and RV^{Mean}, which is the standard simple combination estimator. The results of these tests are presented in Table 4.

The results of the Romano–Wolf test reveal some interesting patterns. Firstly, at the 10% level of significance, *every* estimator significantly outperforms $\text{RV}^{1\text{day}}$, in the full sample and in all three sub-samples. This is clearly a strong signal that high frequency data, when used in any one of the 34 other estimators in this study, yields more precise estimates of QV than daily data can. When $\text{RV}^{5\,\text{min}}$ is taken as a benchmark we see more variation in the results: some estimators are significantly better, others are significantly worse, and some are not significantly different. Broadly stated, the estimators that out-perform $\text{RV}^{5\,\text{min}}$ include RV sampled at the 1-min frequency (either in tick time or calendar time), $\text{RV}^{AC1,1\,\text{min}}$, RK with the Bartlett or TH kernel (when sampled every tick), and RRV, QRV and MedRV, as well as all three combination estimators. The estimators that *under*-perform $\text{RV}^{5\,\text{min}}$ include RV sampled at the 1 h or 1 day frequency (either in tick time or calendar time), $\text{RV}^{BR,bc}$, RK with the cubic, TH or NFP kernel (when sampled at the one-minute frequency), and BPV when sampled at the 5-min frequency.

Finally, when RV^{Mean} is taken as the benchmark estimator in the Romano–Wolf testing method a very clear conclusion emerges: only two estimators significantly out-perform RV^{Mean} in the full sample, namely $\text{RV}^{1\,\text{min}}$ and $\text{RV}^{AC1,1\,\text{min}}$, and *no* individual estimator significantly out-performs RV^{Mean} in any of the three sub-samples. A few estimators are not significantly different, and most individual estimators

[8] The MCS is implemented via a bootstrap re-sampling scheme. We use Politis and Romano's (1994) stationary bootstrap with an average block length of 10 days and 1000 bootstrap replications for each test.

[9] Under the MSE distance, the MCS contains 11 estimators: $\text{RV}^{1\,\text{min}}$, $\text{RV}^{tick,1\,\text{min}}$, $\text{RV}^{AC1,1\,\text{min}}$, TSRV^{1tick}, MSRV^{1tick}, RK^{bart}, $\text{BPV}^{1\,\text{min}}$, QRV, MedRV, MinRV, and $\text{RV}^{Geo\text{-}mean}$. The difference in the number of estimators in the MCS under these two distance measures reflects the power to distinguish between competing estimators.

[10] As noted by a referee, the MCS could be used to form an optimal "trimmed" combination estimator, where only those estimators that are contained in the MCS are included in the combination estimator. Such an estimator will certainly perform well in the sample period used to construct the MCS, and an out-of-sample analysis could be used to determine whether it also performs well on a different sample.

[11] The Romano–Wolf testing method is also implemented using Politis and Romano's (1994) stationary bootstrap with an average block length of 10 days, and we again use 1000 bootstrap replications for each test.

Table 4
Romano–Wolf tests on the realised measures.

	RV^{1day}				$RV^{5\,min}$				RV^{Mean}			
	Full	96–99	00–03	04–08	Full	96–99	00–03	04–08	Full	96–99	00–03	04–08
$RV^{1\,s}$	✓	✓	✓	✓	–	–	–	✓	×	–	×	×
$RV^{5\,s}$	✓	✓	✓	✓	–	–	–	✓	×	–	×	–
$RV^{1\,min}$	✓	✓	✓	✓	✓	✓	–	✓	✓	–	–	–
$RV^{5\,min}$	✓	✓	✓	✓	★	★	★	★	×	×	×	×
RV^{1h}	✓	✓	✓	✓	×	×	×	×	×	×	×	×
RV^{1day}	★	★	★	★	×	×	×	×	×	×	×	×
$RV^{tick,5\,s}$	✓	✓	✓	✓	–	–	–	✓	×	–	×	–
$RV^{tick,1\,min}$	✓	✓	✓	✓	✓	✓	–	✓	–	–	–	–
$RV^{tick,5\,min}$	✓	✓	✓	✓	–	–	–	–	×	×	×	×
$RV^{tick,1h}$	✓	✓	✓	✓	×	×	×	×	×	×	×	×
RV^{BR}	✓	✓	✓	✓	–	×	–	✓	×	×	×	×
$RV^{BR,bc}$	✓	✓	✓	✓	×	×	–	×	×	×	×	×
$RV^{AC1,1\,min}$	✓	✓	✓	✓	✓	✓	–	✓	✓	–	–	–
$RV^{AC1,5\,min}$	✓	✓	✓	✓	–	–	–	–	×	×	×	×
$TSRV^{tick}$	✓	✓	✓	✓	–	×	–	✓	×	×	×	×
$TSRV^{tick,1\,min}$	✓	✓	✓	✓	–	×	–	✓	×	×	×	×
$MSRV^{tick}$	✓	✓	✓	✓	–	×	–	✓	×	×	×	×
$MSRV^{tick,1\,min}$	✓	✓	✓	✓	–	×	–	✓	×	×	×	×
RK^{bart}	✓	✓	✓	✓	✓	–	–	✓	×	×	×	–
$RK^{bart,1\,min}$	✓	✓	✓	✓	–	×	–	–	×	×	×	×
RK^{cubic}	✓	✓	✓	✓	–	×	–	✓	×	×	×	×
$RK^{cubic,1\,min}$	✓	✓	✓	✓	×	×	×	×	×	×	×	×
RK^{TH2}	✓	✓	✓	✓	✓	–	–	✓	×	×	×	–
$RK^{TH2,1\,min}$	✓	✓	✓	✓	×	×	–	×	×	×	×	×
RK^{NFP}	✓	✓	✓	✓	–	×	–	✓	×	×	×	×
$RK^{NFP,1\,min}$	✓	✓	✓	✓	×	×	×	×	×	×	×	×
RRV	✓	✓	✓	✓	✓	–	✓	✓	×	×	×	×
$BPV^{1\,min}$	✓	✓	✓	✓	×	×	–	✓	×	×	×	–
$BPV^{5\,min}$	✓	✓	✓	✓	×	×	×	×	×	×	×	×
QRV	✓	✓	✓	✓	✓	–	–	✓	–	–	–	–
MedRV	✓	✓	✓	✓	✓	–	–	✓	×	–	–	×
MinRV	✓	✓	✓	✓	–	–	–	✓	×	×	×	×
RV^{Mean}	✓	✓	✓	✓	✓	✓	✓	✓	★	★	★	★
$RV^{Geo\text{-}mean}$	✓	✓	✓	✓	✓	–	–	✓	×	×	×	×
RV^{Median}	✓	✓	✓	✓	✓	–	–	✓	×	×	×	–

Notes: This table presents the results of the Romano–Wolf "stepwise" test for three different choices of benchmark estimator. Columns 1–4 present indicators for when the benchmark is set to RV^{1day}: using the QLIKE distance, the indicator is set to ✓if the estimator is significantly more accurate than RV^{1day}, to × if the estimator is significantly *less* accurate than RV^{1day}, and to – if the estimator's accuracy is not significantly different from RV^{1day}. The four columns refer to the full sample period and three sub-samples, 1996–1999, 2000–2003 and 2004–2008. Columns 5–8 present corresponding results when the benchmark estimator is set to $RV^{5\,min}$, and columns 9–12 present results when the benchmark estimator is set to RV^{Mean}. The benchmark estimator in each column is indicated with a ★.

are significantly worse.[12] This is a strong endorsement of using this simple combination estimator in practice.

3.4. Results using optimal combination estimators

In this section we present our estimated optimal combination estimators. We consider a standard

[12] When using the MSE distance (see Table 4A in the web appendix), there are fewer significant results: all estimators are still found to significantly out-perform RV^{1day}, but in the comparisons with $RV^{5\,min}$ or RV^{Mean} as the benchmark there are few rejections of the null hypothesis.

parametric combination estimator, namely a linear combination:

$$\hat{X}_t^* = \hat{w}_0 + \sum_{i=1}^{n} \hat{w}_i X_{it}, \tag{11}$$

estimated using Proposition 1 above. We consider both unconstrained combination estimators, which satisfy the condition that the unknown weights all lie in the interior of the parameter space, thus permitting us to compute standard errors using Proposition 1, and constrained combination estimators, with the constraint being that all weights must be non-negative. The constrained estimation acts as a model selection procedure, and makes obtaining standard errors difficult (which is why we do not pursue this here), but provides additional information on the individual estimators that are most useful in a combination estimator.

Table 5 presents the results for optimal linear combinations under the QLIKE distance. In the optimal constrained combination estimator we see substantial weight on both QRV and $\text{RV}^{tick,5\,\text{min}}$, and we also see non-zero weights on a collection of simple RVs, with sampling frequencies from 5 s up to and including 1 day. The unconstrained combination has few individually significant coefficients, though there are significant coefficients on BPV sampled at the 1-min frequency, $\text{RV}^{BR,bc}$ and $\text{RV}^{AC1,1\,\text{min}}$, and simple RV with sampling frequencies of 1 h and 1 day.

Often of interest in the forecasting literature is the question of whether the estimated optimal combination is significantly different from a simple equally-weighted average. If we let w_i^* denote the optimal linear combination weights, the hypotheses of interest are:

$$H_0 : w_0^* = 0 \cap w_1^* = w_2^* = \cdots = w_n^* = 1/n \tag{12}$$
$$\text{vs. } H_a : w_0^* \neq 0 \cup w_i^* \neq 1/n \quad \text{for some } i = 1, 2, \ldots, n.$$

Using Proposition 1 these hypotheses can be tested using Wald tests, and we find, for the full sample of data, that the null that the optimal combination is an equally-weighted combination can be rejected with a p-value of less than 0.001, under both the MSE and the QLIKE distance. Thus, while Section 3.3 revealed that the simple mean was not consistently beaten by any individual estimator, it can still be improved: an optimally formed linear combination is significantly more accurate than an equally-weighted average.

Finally, we conduct a set of tests related to idea of forecast encompassing; see Chong and Hendry (1986) and Fair and Shiller (1990). We test the null hypothesis that a single realised measure (i) encompasses the information in all other estimators:

$$H_0^i : w_i^* = 1 \cap w_j^* = 0 \quad \forall\, j \neq i \tag{13}$$
$$\text{vs. } H_a^i : w_i^* \neq 1 \cup w_j^* \neq 0 \quad \text{for some } j \neq i,$$

$i = 1, 2, \ldots, n$. We find that the null hypothesis is rejected for every single estimator, under both the MSE and the QLIKE distance, with all p-values being less than 0.001. This is strong evidence that there are gains from considering combination estimators of quadratic variation: no single estimator dominates all others. This result is new to the realised volatility literature, but is probably not surprising to those familiar with forecasting in practice.

3.5. *Results from an out-of-sample forecasting experiment*

Our results above suggest that there are gains from using combination estimators of the volatility of IBM stock returns, in terms of average accuracy. In this section we study whether these gains in estimation accuracy translate into gains in *forecast* accuracy. We do this via a simple out-of-sample forecasting experiment. We use each of the individual estimators, as well as the combination estimators, in a heterogeneous autoregressive (HAR) model (see Corsi, 2004, and Müller et al., 1997), which has been shown to work well in volatility forecasting problems, see Andersen et al. (2007), for example. This model is designed to capture some of the long memory-type properties of volatility in a simple autoregressive framework, by using estimates of volatility over the past day, week (5 trading days) and month (22 trading days) as predictors of future volatility. The model is specified as:

$$\tilde{\theta}_t = \beta_{0i} + \beta_{Di} X_{it-1} + \beta_{Wi} \frac{1}{5} \sum_{j=1}^{5} X_{i,t-j} + \beta_{Mi} \frac{1}{22} \sum_{j=1}^{22} X_{i,t-j} + \varepsilon_{it}. \tag{14}$$

Table 5
Optimal linear combination weights.

	Full	96–99	00–03	04–08	Full	96–99	00–03	04–08
Constant	**0.181**	**1.084**	**0.181**	**0.267**	0.154	0.994	0.154	0.223
$RV^{1\text{ s}}$	0.023	0.269	−0.020	0.865	0.031	0.004	0.030	0.000
$RV^{5\text{ s}}$	0.154	−0.499	0.119	0.517	0.080	0	0.112	0.079
$RV^{1\text{ min}}$	0.408	−0.274	−0.169	0.765	0.036	0	0	0.067
$RV^{5\text{ min}}$	0.168	0.045	−0.103	**0.927**	0.054	0	0	0.032
$RV^{1\text{h}}$	**−0.075**	−0.071	**−0.101**	−0.025	0	0	0	0
$RV^{1\text{day}}$	**0.049**	**0.088**	**0.028**	**0.034**	0.048	0.074	0.005	0.037
$RV^{tick,5\text{ s}}$	0.044	0.200	0.061	**−1.223**	0.105	0.016	0.123	0
$RV^{tick,1\text{ min}}$	0.081	−0.312	**0.554**	−0.169	0.064	0	0.148	0
$RV^{tick,5\text{ min}}$	0.082	**0.463**	−0.122	0.170	0.150	0.196	0	0
$RV^{tick,1\text{h}}$	0.037	0.008	0.002	0.055	0	0	0	0.017
RV^{BR}	0.077	0.247	−0.049	**0.366**	0	0.039	0	0.262
$RV^{BR,bc}$	**−0.229**	−0.006	−0.189	−0.124	0	0	0	0
$RV^{AC1,1\text{ min}}$	**0.408**	**0.714**	**0.604**	**−0.714**	0.096	0.192	0.080	0
$RV^{AC1,5\text{ min}}$	−0.004	−0.101	0.035	−0.264	0.008	0	0	0
$TSRV^{tick}$	0.069	−0.060	0.043	−0.033	0	0	0	0.042
$TSRV^{tick,1\text{ min}}$	0.252	0.380	**0.463**	0.292	0	0.052	0.119	0.022
$MSRV^{tick}$	−0.026	0.192	−0.234	0.099	0	0.002	0	0.070
$MSRV^{tick,1\text{ min}}$	−0.103	0.191	−0.507	−0.469	0	0	0	0
RK^{bart}	−0.034	−0.119	0.107	0.473	0	0	0	0
$RK^{bart,1\text{ min}}$	0.023	−0.298	0.151	0.477	0	0	0.051	0
RK^{cubic}	0.079	0.025	−0.122	0.179	0	0	0	0
$RK^{cubic,1\text{ min}}$	0.020	−0.220	0.273	0.116	0	0	0.064	0
RK^{TH2}	−0.144	−0.460	−0.021	−0.229	0	0	0	0
$RK^{TH2,1\text{ min}}$	0.011	0.100	0.037	−0.443	0.013	0	0	0
RK^{NFP}	−0.014	−0.387	0.038	−0.319	0	0	0	0
$RK^{NFP,1\text{ min}}$	−0.061	0.479	−0.196	−0.316	0	0	0	0
RRV	−0.173	0.300	−0.292	**−1.163**	0	0	0	0
$BPV^{1\text{ min}}$	**−0.535**	0.314	−0.430	1.380	0	0.005	0	0.062
$BPV^{5\text{ min}}$	−0.081	**−0.314**	**0.357**	−0.060	0	0	0.128	0
QRV	0.056	**0.266**	0.260	0.127	0.192	0.073	0	0.046
MedRV	0.234	−0.427	0.364	0.424	0	0	0	0
MinRV	−0.098	−0.094	−0.159	**−1.111**	0	0	0	0

Notes: This table presents the QLIKE optimal linear combination weights. Columns 1–4 present these weights for the unconstrained case, with estimates that are significantly different from zero at the 0.10 level highlighted in bold. (Standard errors were computed using Proposition 1, with Newey & West, 1987, estimates of the covariance matrix, B_T, but are not reported here in the interests of space.) The four columns refer to the full sample period and three sub-samples, 1996–1999, 2000–2003 and 2004–2008. Columns 5–8 present the optimal linear combination weights, using the QLIKE distance, imposing the constraint that each weight must be weakly positive. Weights that were on the boundary at zero are reported as "0", while those away from the boundary are reported to three decimal places.

We estimate this model for each of our 32 individual estimators, and our 3 simple combination estimators. We use the period from January 1996 to December 1999 (1011 trading days) as the initial estimation period, and re-estimate the model each day using a rolling window of 1011 days of data. For each day, we construct a one-step-ahead forecast using the estimated parameters for that day,[13] thus replicating the real-time forecasting problem faced in practice. With the volatility forecasts based on each of the 35 estimators, and an unbiased proxy for the true

[13] For simplicity we focus on one-step-ahead forecasts, and leave the interesting extension to multi-step-ahead forecasts for future work.

latent quadratic variation, we can then conduct comparison tests of forecast accuracy similar to those used in the previous sections for comparing *estimation* accuracy.

In Table 6 we present the first results from our out-of-sample forecasting experiment. These tables present the average difference in QLIKE distance, relative to the forecast based on $\mathrm{RV}^{5\,\mathrm{min}}$, as well as the ranks of the various forecasts for the full sample and for two sub-samples, 2000–2003 and 2004–2008. (The first sub-sample, 1996–1999, is used for the initial estimation of the models, and thus cannot be used in the comparison of forecast accuracy.) In addition to the 32 individual realised measures and the 3 simple combination estimators, we also consider 3 combination *forecasts*, based on an equally-weighted arithmetic mean of the forecasts, an equally-weighted geometric mean of the forecasts, and the median of the forecasts, leading to a total of 38 different forecasts.[14]

Table 6 shows that the best forecast overall is that based on $\mathrm{RV}^{AC1,1\,\mathrm{min}}$, followed by $\mathrm{RV}^{Geo\text{-}mean}$ and $\mathrm{RV}^{1\,\mathrm{min}}$. It is interesting to note that $\mathrm{RV}^{AC1,1\,\mathrm{min}}$ performs quite a bit worse in the second sub-sample than in the first sub-sample, while $\mathrm{RV}^{Geo\text{-}mean}$ displays the opposite behaviour. The deterioration in the performance of $\mathrm{RV}^{AC1,1\,\mathrm{min}}$ in the second sub-sample may be related to the increased amount of trading in this sub-sample, which presumably reduces the impact of market microstructure effects at the one-minute frequency. The simple $\mathrm{RV}^{1\,\mathrm{min}}$ and $BPV^{5\,\mathrm{min}}$ estimators both exhibit good and stable performance over the two sub-samples. We again estimate the model confidence set of Hansen et al. (2005), and find that it contains just 7 out of the 38 total forecasts for the full sample. Interestingly, it contains two forecasts based on combination *estimators* (RV^{Mean} and $\mathrm{RV}^{Geo\text{-}mean}$), but none of the combination *forecasts*. Also noteworthy is the fact that in the 2004–2008 sub-sample, the MCS contains just a single forecast, namely that based on $\mathrm{RV}^{Geo\text{-}mean}$. Overall, the results in Table 6 again support the use of combination estimators of volatility. The rankings are similar to those based directly on estimation accuracy in Table 3, though they are not identical: QRV, for example, does worse in terms of forecast accuracy than in terms of estimation accuracy, while $BPV^{5\,\mathrm{min}}$ does better.

In Tables 7 and 8 we present the results of Romano and Wolf (2005) tests for determining the forecasts that are significantly better or worse than a given benchmark forecast. As in the previous section, we initially consider three choices of benchmark forecasts: those based on $\mathrm{RV}^{1\mathrm{day}}$, $\mathrm{RV}^{5\,\mathrm{min}}$ and RV^{Mean}. Table 7 shows that every forecast significantly out-performs $\mathrm{RV}^{1\mathrm{day}}$. The results with $\mathrm{RV}^{5\,\mathrm{min}}$ show more variation: several forecasts are found to be significantly worse than $\mathrm{RV}^{5\,\mathrm{min}}$, several are not significantly different, and a number are significantly better. The better forecasts include $\mathrm{RV}^{1\,\mathrm{min}}$, RV^{BR}, $\mathrm{RV}^{AC1,1\,\mathrm{min}}$, RRV, $BPV^{5\,\mathrm{min}}$ and all 6 combinations (3 based on combination estimators, and 3 based on combinations of forecasts). With RV^{Mean} as the benchmark (see Table 8), we find that most forecasts are significantly worse, and *none* are significantly better in the full sample. These tables thus provide further evidence in support of the use of combination estimators of volatility.

In Table 8 we also conduct Romano–Wolf tests, with the equally-weighted arithmetic mean of the forecasts, FCAST^{Mean}, as the benchmark. We find that most individual forecasts are significantly worse, and a few are significantly better.[15] These better forecasts include $\mathrm{RV}^{1\,\mathrm{min}}$, RV^{BR}, $\mathrm{RV}^{AC1,1\,\mathrm{min}}$ and RRV, as well as all three forecasts based on combination estimators, RV^{Mean}, $\mathrm{RV}^{Geo\text{-}mean}$ and RV^{Median}. This reveals that significantly better forecasts could be obtained by building a single forecast based on a combination estimator, rather than by combining many forecasts based on individual estimators. This is a novel finding. We investigate it further by conducting simple Diebold and Mariano (1995) tests of equal predictive accuracy for the pair-wise comparison of RV^i with FCAST^i, where $i = \{Mean, Geo\text{-}mean, Median\}$. These tests allow us to more directly answer the question of whether it is better to combine estimators and then forecast, or to combine forecasts based on many estimators. In these tests, positive (negative) t-statistics indicate that FCAST^i had a larger (smaller) average forecast error than RV^i, and t-statistics greater than 1.64 in absolute

[14] Table 6A of the web appendix presents corresponding results for differences in MSE distances for these forecasts.

[15] Under the MSE, we find little evidence that any forecast is significantly different, see Table 8A in the web appendix.

Table 6
Performance of the out-of-sample forecasts.

	Avg. ΔQLIKE	Rank			In MCS?		
		Full	00–03	04–08	Full	00–03	04–08
$RV^{1\,s}$	0.006	33	29	33	–	–	–
$RV^{5\,s}$	0.001	28	17	30	–	–	–
$RV^{1\,min}$	−0.012	3	5	6	✓	–	–
$RV^{5\,min}$	0	18	20	19	–	–	–
RV^{1h}	0.024	36	36	35	–	–	–
RV^{1day}	0.070	38	38	38	–	–	–
$RV^{tick,5\,s}$	0.006	34	30	34	–	–	–
$RV^{tick,1\,min}$	−0.010	8	7	12	–	–	–
$RV^{tick,5\,min}$	0.001	21	25	21	–	–	–
$RV^{tick,1h}$	0.022	35	35	36	–	–	–
RV^{BR}	−0.010	9	13	3	✓	–	–
$RV^{BR,bc}$	0.040	37	37	37	–	–	–
$RV^{AC1,1\,min}$	−0.013	1	3	9	✓	✓	–
$TSRV^{AC1,5\,min}$	0.001	25	24	23	–	–	–
$TSRV^{tick}$	−0.002	17	19	13	–	–	–
$MSRV^{tick,1\,min}$	0.000	19	22	18	–	–	–
$MSRV^{tick}$	0.001	23	31	14	–	–	–
$RV^{tick,1\,min}$	0.004	31	34	29	–	–	–
RK^{bart}	0.001	27	28	17	–	–	–
$RK^{bart,1\,min}$	0.001	20	18	24	–	–	–
RK^{cubic}	0.001	22	27	16	–	–	–
$RK^{cubic,1\,min}$	0.004	32	32	32	–	–	–
RK^{TH2}	0.003	30	33	25	–	–	–
$RK^{TH2,1\,min}$	0.002	29	23	27	–	–	–
RK^{NFP}	0.001	24	26	20	–	–	–
$RK^{NFP,1\,min}$	0.001	26	21	26	–	–	–
RRV	−0.011	7	9	7	✓	–	–
$BPV^{1\,min}$	−0.007	10	6	15	–	–	–
$BPV^{5\,min}$	−0.011	5	8	5	✓	–	–
QRV	−0.007	13	1	31	–	✓	–
MedRV	−0.007	12	2	28	–	✓	–
MinRV	−0.007	11	4	22	–	–	–
RV^{Mean}	−0.012	4	10	2	✓	–	–
$RV^{Geo\text{-}mean}$	−0.013	2	11	1	✓	–	✓
RV^{Median}	−0.011	6	12	4	–	–	–
$FCAST^{Mean}$	−0.006	15	15	10	–	–	–
$FCAST^{Geo\text{-}mean}$	−0.007	14	14	8	–	–	–
$FCAST^{Median}$	−0.005	16	16	11	–	–	–

Notes: The first column of this table presents the average difference in QLIKE of the forecast based on each realised measure relative to the forecast based on $RV^{5\,min}$, with negative (positive) values indicating that the forecast was on average closer to (further from) to the target variable than $RV^{5\,min}$. Columns 2–4 present the rank of each forecast using QLIKE, for the full sample period and for two sub-samples, 2000–2003 and 2004–2008. The most accurate forecast is ranked 1, and the least accurate is ranked 38. Columns 5–7 present an indicator for whether the estimator was in the "model confidence set" at the 90% confidence level (equal to ✓ if in, – if not) in the full sample and each of the two sub-samples.

value indicate that the difference is significant at the 0.10 level. We found that the t-statistics for the full-sample were 7.66, 7.10 and 6.49 for the *Mean*, *Geo-mean* and *Median*, and that the corresponding t-statistics in the sub-samples were all greater than 3.50. Thus, we find strong evidence that forecasts based

Table 7
Romano–Wolf tests on forecasts, with RV^{1day} and $RV^{5\,min}$ as benchmarks.

	RV^{1day}			$RV^{5\,min}$		
	Full	00–03	04–08	Full	00–03	04–08
$RV^{1\,s}$	✓	✓	✓	–	–	–
$RV^{5\,s}$	✓	✓	✓	–	–	–
$RV^{1\,min}$	✓	✓	✓	✓	✓	–
$RV^{5\,min}$	✓	✓	✓	★	★	★
RV^{1h}	✓	✓	✓	×	×	–
RV^{1day}	★	★	★	×	×	×
$RV^{tick,5\,s}$	✓	✓	✓	–	–	–
$RV^{tick,1\,min}$	✓	✓	✓	✓	✓	–
$RV^{tick,5\,min}$	✓	✓	✓	–	–	–
$RV^{tick,1h}$	✓	✓	✓	×	×	×
RV^{BR}	✓	✓	✓	✓	✓	✓
$RV^{BR,bc}$	✓	✓	✓	×	×	×
$RV^{AC1,1\,min}$	✓	✓	✓	✓	✓	–
$RV^{AC1,5\,min}$	✓	✓	✓	×	–	×
$TSRV^{tick}$	✓	✓	✓	–	–	–
$TSRV^{tick,1\,min}$	✓	✓	✓	–	–	–
$MSRV^{tick}$	✓	✓	✓	–	–	–
$MSRV^{tick,1\,min}$	✓	✓	✓	×	×	–
RK^{bart}	✓	✓	✓	–	–	–
$RK^{bart,1\,min}$	✓	✓	✓	–	–	–
RK^{cubic}	✓	✓	✓	–	–	–
$RK^{cubic,1\,min}$	✓	✓	✓	×	–	×
RK^{TH2}	✓	✓	✓	–	–	–
$RK^{TH2,1\,min}$	✓	✓	✓	–	–	–
RK^{NFP}	✓	✓	✓	–	–	–
$RK^{NFP,1\,min}$	✓	✓	✓	–	–	–
RRV	✓	✓	✓	✓	✓	✓
$BPV^{1\,min}$	✓	✓	✓	–	✓	–
$BPV^{5\,min}$	✓	✓	✓	✓	✓	✓
QRV	✓	✓	✓	–	✓	–
MedRV	✓	✓	✓	–	✓	–
MinRV	✓	✓	✓	–	✓	–
RV^{Mean}	✓	✓	✓	✓	✓	✓
$RV^{Geo\text{-}mean}$	✓	✓	✓	✓	✓	✓
RV^{Median}	✓	✓	✓	✓	✓	✓
$FCAST^{Mean}$	✓	✓	✓	✓	–	✓
$FCAST^{Geo\text{-}mean}$	✓	✓	✓	✓	–	✓
$FCAST^{Median}$	✓	✓	✓	✓	–	✓

Notes: This table presents the results of the Romano–Wolf "stepwise" test for two different choices of benchmark forecast. Columns 1–3 present indicators for when the benchmark is set to forecasts based on RV^{1day}: using the QLIKE distance, the indicator is set to ✓if the forecast is significantly more accurate than RV^{1day}, to × if the forecast is significantly less accurate than RV^{1day}, and to – if the forecast's accuracy is not significantly different from RV^{1day}. The three columns refer to the full sample period and two sub-samples, 2000–2003 and 2004–2008. Columns 4–6 present corresponding results when the benchmark forecast is based on $RV^{5\,min}$. The benchmark forecast in each column is indicated by a ★.

on combination RV estimators are significantly better than the corresponding combination forecasts.[16]

[16] Under the MSE the results were mixed; in no case do we find that the differences are significant.

Overall, this study of out-of-sample forecast accuracy yields two main conclusions. Firstly (and unsurprisingly), better estimation accuracy generally leads to better forecast accuracy. The rankings based

Table 8
Romano–Wolf tests on forecasts, with RV^{Mean} and $FCAST^{Mean}$ as benchmarks.

	RV^{Mean}			$FCAST^{Mean}$		
	Full	00–03	04–08	Full	00–03	04–08
$RV^{1\text{ s}}$	×	×	×	×	–	×
$RV^{5\text{ s}}$	×	×	×	×	–	×
$RV^{1\text{ min}}$	–	–	–	✓	✓	–
$RV^{5\text{ min}}$	×	×	×	×	–	×
$RV^{1\text{h}}$	×	×	×	×	×	×
$RV^{1\text{day}}$	×	×	×	×	×	×
$RV^{tick,5\text{ s}}$	×	×	×	×	–	×
$RV^{tick,1\text{ min}}$	–	–	×	✓	✓	–
$RV^{tick,5\text{ min}}$	×	×	×	×	–	×
$RV^{tick,1\text{h}}$	×	×	×	×	×	×
RV^{BR}	–	–	–	✓	–	–
$RV^{BR,bc}$	×	×	×	×	×	×
$RV^{AC1,1\text{ min}}$	–	✓	–	✓	✓	–
$RV^{AC1,5\text{ min}}$	×	×	×	×	–	×
$TSRV^{tick}$	×	×	×	–	–	–
$TSRV^{tick,1\text{ min}}$	×	×	×	×	×	×
$MSRV^{tick}$	×	×	×	×	×	×
$MSRV^{tick,1\text{ min}}$	×	×	×	×	×	×
RK^{bart}	×	×	×	×	×	×
$RK^{bart,1\text{ min}}$	×	×	×	×	–	×
RK^{cubic}	×	×	×	×	×	×
$RK^{cubic,1\text{ min}}$	×	×	×	×	×	×
RK^{TH2}	×	×	×	×	×	×
$RK^{TH2,1\text{ min}}$	×	×	×	×	–	×
RK^{NFP}	×	×	×	×	×	×
$RK^{NFP,1\text{ min}}$	×	×	×	×	–	×
RRV	–	–	×	✓	✓	–
$BPV^{1\text{ min}}$	–	–	×	–	✓	–
$BPV^{5\text{ min}}$	–	–	–	–	✓	–
QRV	–	–	×	–	✓	–
MedRV	–	–	×	–	✓	–
MinRV	–	–	×	–	✓	–
RV^{Mean}	★	★	★	✓	✓	✓
$RV^{Geo\text{-}mean}$	–	–	✓	✓	✓	✓
RV^{Median}	–	–	×	✓	✓	–
$FCAST^{Mean}$	×	×	×	★	★	★
$FCAST^{Geo\text{-}mean}$	×	×	×	✓	✓	–
$FCAST^{Median}$	×	×	×	–	–	×

Notes: This table presents the results of the Romano–Wolf "stepwise" test for two different choices of benchmark forecast. Columns 1–3 present indicators for when the benchmark is set to forecasts based on RV^{Mean}: using the QLIKE distance, the indicator is set to ✓if the forecast is significantly more accurate than RV^{Mean}, to × if the forecast is significantly less accurate than RV^{Mean}, and to – if the forecast's accuracy is not significantly different from $RV^{1\text{day}}$. The three columns refer to the full sample period and two sub-samples, 2000–2003 and 2004–2008. Columns 4–6 present corresponding results when the benchmark forecast is based on $FCAST^{Mean}$. The benchmark forecast in each column is indicated by a ★.

on estimation accuracy are not identical to those based on forecast accuracy, however, with the move to forecasting benefiting some estimators but not others. Secondly, forecasts based on combinations of RV estimators significantly out-perform combinations of forecasts based on individual RV estimators.

4. Summary and conclusion

Recent advances in financial econometrics have led to the development of new estimators of asset price variability using high frequency price data. These estimators are based on a variety of different assumptions and take many different functional forms. Motivated by the success of combination forecasts over individual forecasts in a range of forecasting applications (see Clemen, 1989, and Timmermann, 2006, for example), this paper sought to answer the question: do combinations of individual estimators offer accuracy gains relative to individual estimators? The answer is a resounding "yes".

This paper presents a novel method for combining individual realised measures to form new estimators of price variability, overcoming the obstacle that the quantity of interest is not observable, even ex post. We applied this method to a collection of 32 different realised measures, across 8 distinct classes of estimators, estimated on high frequency IBM price data over the period 1996–2008. Using the Romano and Wolf (2005) test, we find that, in terms of average accuracy, only two individual realised measures significantly out-perform a simple equally-weighted average. In terms of out-of-sample forecast accuracy, with forecasts based on a simple HAR model, we find that *no* individual estimator can significantly out-perform a forecast based on a simple equally-weighted average. Further, we find that none of the individual estimators encompasses the information in all other estimators, providing further support for the use of combination realised measures. Overall, our results suggest that there are indeed benefits from combining the information contained in the array of different volatility estimators proposed in the literature to date.

Acknowledgements

We thank Asger Lunde, Neil Shephard, an anonymous referee, and seminar participants at the third ESRC seminar on Nonlinear Economics and Finance at Keele University.

References

Aït-Sahalia, Y., Mykland, P., & 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. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*, 885–905.

Andersen, T. G., Bollerslev, T., & Meddahi, N. (2005). Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica*, *73*, 279–296.

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. (2001a). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, *96*, 42–55.

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

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*, 579–626.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., & Diebold, F. X. (2006). Volatility and correlation forecasting. In G. Elliott, et al., (Eds.), *Handbook of economic forecasting*. Amsterdam: North Holland Press.

Andersen, T. G., Dobrev, D., & Schaumburg, E. (2008). *Jump robust volatility estimation*. Working paper, Northwestern University.

Bandi, F. M., & Russell, J. R. (2006). Separating microstructure noise from volatility. *Journal of Financial Economics*, *79*, 655–692.

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., & 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). Econometric analysis of realized covariation: High frequency based covariance, regression and correlation in financial economics. *Econometrica*, *72*, 885–925.

Barndorff-Nielsen, O. E., & Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, *4*, 1–30.

Barndorff-Nielsen, O. E., & Shephard, N. (2007). Variation, jumps, market frictions and high frequency data in financial econometrics. In R. Blundell, W. K. Newey, & T. Persson (Eds.), *Advances in economics and econometrics. Theory and applications*: *Vol. 3*. *Econometric Society Monographs* (pp. 328–372). Cambridge University Press.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*, *76*(6), 1481–1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Realized kernels in practice: Trades and quotes. *Econometrics Journal* (in press).

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operations Research Quarterly*, *20*, 451–468.

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

Chong, Y. Y., & Hendry, D. F. (1986). Econometric evaluation of linear macroeconomic models. *Review of Economic Studies*, *53*, 671–690.

Christensen, K., & Podolskij, M. (2007). Realized range-based estimation of integrated variance. *Journal of Econometrics*, *141*, 323–349.

Christensen, K., Oomen, R.C.A., & Podolskij, M. (2008). *Realised quantile-based estimation of the integrated variance*. Working paper.

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

Clements, M. P., & Hendry, D. F. (1998). *Forecasting economic time series*. Cambridge: Cambridge University Press.

Corsi, F. (2004). *A simple long memory model of realized volatility*. Working paper. University of Southern Switzerland.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. Oxford: Oxford University Press.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.

Fair, R. C., & Shiller, R. J. (1990). Comparing information in forecasts from econometric models. *American Economic Review*, *80*, 375–389.

French, K. R., Schwert, G. W., & Stambaugh, R. F. (1987). Expected stock returns and volatility. *Journal of Financial Economics*, *19*, 3–29.

Gallant, A. R., & White, H. (1988). *A unified theory of estimation and inference for nonlinear dynamic models*. New York: Basil Blackwell.

Halperin, M. (1961). Almost linearly-optimum combination of unbiased estimates. *Journal of the American Statistical Association*, *56*, 36–43.

Hansen, P. R., & Lunde, A. (2006a). Realized variance and market microstructure noise. *Journal of Business and Economic Statistics*, *24*, 127–161.

Hansen, P. R., & Lunde, A. (2006b). Consistent ranking of volatility models. *Journal of Econometrics*, *131*, 97–121.

Hansen, P. R., Lunde, A., & Nason, J. M. (2005). *Model confidence sets for forecasting models*. Federal Reserve Bank of Atlanta Working Paper 2005-7.

Large, J. (2005). *Estimating quadratic variation when quoted prices change by a constant increment*. Working paper. Department of Economics, University of Oxford.

Martens, M., & van Dijk, D. (2007). Measuring volatility with the realized range. *Journal of Econometrics*, *138*, 181–207.

Meddahi, N. (2001). *A theoretical comparison between integrated and realized volatilities*. Manuscript. Université de Montréal.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, *8*, 323–361.

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Puctet, O. V., & von Weizsäcker, J. (1997). Volatilities of different time resolutions – Analyzing the dynamics of market components. *Journal of Empirical Finance*, *4*, 213–239.

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, *55*, 299–306.

Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts (with discussion). *Journal of the Royal Statistical Society, Series A*, *137*, 131–149.

Newey, W. K., & West, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703–708.

Oomen, R. C. A. (2006). Properties of realized variance under alternative sampling schemes. *Journal of Business and Economic Statistics*, *24*, 219–237.

Patton, A. J. (2006). *Volatility forecast comparison using imperfect volatility proxies*. Research Paper 175. Quantitative Finance Research Centre, University of Technology Sydney.

Patton, A. J. (2008). *Data-based ranking of realised volatility estimators*. Working paper. Oxford-Man Institute of Quantitative Finance, University of Oxford.

Patton, A. J., & Sheppard, K. (2008). Evaluating volatility and correlation forecasts. In T. G. Andersen et al., (Eds.), *Handbook of financial time series*. Springer Verlag (in press).

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

Reid, D. J. (1968). Combining three estimates of gross domestic product. *Economica*, *35*, 431–444.

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

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

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. G. Timmermann (Eds.), *Handbook of economic forecasting*. Amsterdam: North Holland Press.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, *12*, 1019–1043.

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.

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business and Economic Statistics*, *14*, 45–52.

Andrew Patton is a Reader in Economics at the University of Oxford and Deputy-Director of the Oxford-Man Institute of Quantitative Finance. He previously taught at the London School of Economics and served as an academic consultant for the Bank of England. His research focuses on financial econometrics, with an emphasis on forecasting risk and dependence, and the analysis of hedge funds.

Kevin Sheppard is a University Lecturer and research associate at Oxford-Man Institute of Quantitative Finance. He is also an official fellow at Keble College Oxford. His research interests include volatility and correlation measurement and modelling, high dimension econometrics and the measurement of uncertainty in economic models.