Journal of Econometrics 187 (2015) 293–311

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes

Lily Y. Liu [a], Andrew J. Patton [a,*], Kevin Sheppard [b]

[a] Duke University, United States

[b] University of Oxford, United States

ARTICLE INFO

Article history:
Received 6 December 2012
Received in revised form
23 August 2014
Accepted 3 February 2015
Available online 9 March 2015

JEL classification:
C58
C22
C53

Keywords:
Realized variance
Volatility forecasting
High frequency data

ABSTRACT

We study the accuracy of a variety of estimators of asset price variation constructed from high-frequency data ("realized measures"), and compare them with a simple "realized variance" (RV) estimator. In total, we consider over 400 different estimators, using 11 years of data on 31 different financial assets spanning five asset classes. When 5-minute RV is taken as the benchmark, we find little evidence that it is outperformed by any other measures. When using inference methods that do not require specifying a benchmark, we find some evidence that more sophisticated measures outperform. Overall, we conclude that it is difficult to significantly beat 5-minute RV.

1. Introduction

In the past fifteen years, many new estimators of asset return volatility constructed using high frequency price data have been developed (see Andersen et al. (2006), Barndorff-Nielsen and Shephard (2007), Meddahi (2011) and Aït-Sahalia and Jacod (2014), *inter alia*, for recent surveys and collections of articles). These estimators generally aim to estimate the quadratic variation or the integrated variance of a price process over some interval of time, such as one day or week. We refer to estimators of this type collectively as "realized measures". This area of research has provided practitioners with an abundance of alternatives, inducing demand for some guidance on which estimators to use in empirical applications. In addition to selecting a particular estimator, these nonparametric measures often require additional choices for their implementation. For example, the practitioner must choose the sampling frequency to use and whether to sample prices in calendar time (every x seconds) or tick-time (every x trades). When both transaction and quotation prices are available, the choice of which price to use also arises. Finally, some realized measures further require choices about tuning parameters such as a kernel bandwidth or "block size".

The aim of this paper is to provide guidance on the choice of realized measure to use in applications. We do so by studying the performance of a large number of realized measures across a broad range of financial assets. In total we consider over 400 realized measures, across eight distinct classes of estimators, and we apply these to 11 years of daily data on 31 individual financial assets covering five asset classes. We compare the realized measures in terms of their estimation accuracy for the latent true quadratic variation, and in terms of their forecast accuracy when combined with a simple and well-known forecasting model. We employ model-free data-based comparison methods that make minimal assumptions on properties of the efficient price process or on the market microstructure noise that contaminates the efficient prices.

To our knowledge, no existing papers have used formal tests to compare the estimation accuracy of a large number of realized measures using real financial data. The fact that the target variable (quadratic variation) is latent, even ex-post, creates an obstacle to applying standard techniques. Previous research on the selection of estimators of quadratic variation has often focused on recommending a sampling frequency based on the underlying theory using plug-in type estimators of nuisance parameters. For some estimators, a formula for the optimal sampling frequency under a set of assumptions is derived and can be computed

* Correspondence to: Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham, NC 27708-0097, United States.
E-mail address: andrew.patton@duke.edu (A.J. Patton).

http://dx.doi.org/10.1016/j.jeconom.2015.02.008
0304-4076/

using estimates of higher order moments, see Bandi and Russell (2008) among others. However, these formulas are usually heavily dependent on assumptions about the microstructure noise and efficient price process, such as independence of the noise from the price and a lack of serial correlation in the noise. Gatheral and Oomen (2010) use simulated data from an agents-based model to evaluate a variety of realized measures, and include recommendations on data sampling and implementing the estimators they study.

Many papers that introduce novel realized measures provide evidence that details the new estimator's advantages over previous estimators. This evidence can be in the form of theoretical properties of estimators such as consistency, asymptotic efficiency, and rate of convergence, or results from Monte Carlo simulations using common stochastic volatility models. These comparisons inevitably require making specific assumptions on important properties of the price process. Empirical applications are also common, although typically only a small number of assets from a single asset class are used, and it is rare that any formal comparison testing is carried out. Moreover, most papers proposing new estimators consider (perhaps reasonably) only a relatively small range of alternative estimators.

Our objective is to compare a large number of available realized measures in a unified, data-based framework. We use the data-based ranking method of Patton (2011a), which makes no assumptions about the properties of the market microstructure noise, aside from standard moment and mixing conditions. The main contribution of this paper is an empirical study of the relative performance of estimators of daily quadratic variation from 8 broad classes of realized measures using data from 31 financial assets spanning different classes. We obtain tick-by-tick transaction and quotation prices from January 2000 to December 2010, and additionally sample prices in calendar-time and tick-time, using sampling frequencies varying from 1 s to 15 min. We use the "model confidence set" of Hansen et al. (2011) to construct sets of realized measures that contain the best measure with a given level of confidence. We are also interested whether a simple RV estimator with a reasonable choice of sampling frequency, namely 5-minute RV, can stand in as a "good enough" estimator for QV, for the assets we consider. This is similar to the comparison of more sophisticated volatility models with a simple benchmark model presented in Hansen and Lunde (2005). We use the stepwise multiple testing method of Romano and Wolf (2005), which allows us to determine whether any of the 400+ competing realized measures is significantly more accurate than a simple realized variance measure based on 5-minute returns. We also conduct an out-of-sample forecasting experiment to study the accuracy of volatility forecasts based on these individual realized measures, when used in the "heterogeneous autoregressive" (HAR) forecasting model of Corsi (2009), for forecast horizons ranging from 1 to 50 trading days.

Finally, we undertake a panel investigation of the market microstructure and market condition variables that explain the differences in the accuracy of the realized measures considered in this paper. While 5-minute RV is beaten on average by (well-chosen) more sophisticated alternatives, the differences are smaller when microstructure noise, somehow measured, is higher, or when volatility is higher. Interestingly, we also find that more sophisticated realized measures generally perform significantly worse in non-US markets than in US markets, the latter having been the focus of much of this literature. This is potentially indicative of different market microstructure effects in non-US markets, which may be better handled with new approaches.

The remainder of this paper is organized as follows. Section 2 provides a brief description of the classes of realized measures. Section 3 describes ranking methodology and tests used to compare the realized measures. Section 4 describes the high frequency data and the set of realized measures we construct. Our main analysis is presented in Section 5, and Section 6 concludes.

2. Measures of asset price variability

To fix ideas and notation, consider a general jump-diffusion model for the log-price p of an asset:

$$dp(t) = \mu(t)\,dt + \sigma(t)\,dW(t) + \kappa(t)\,dN(t) \tag{1}$$

where μ is the instantaneous drift, σ is the (stochastic) volatility, W is a standard Brownian motion, κ is the jump size, and N is a counting measure for the jumps. In the absence of jumps the third term on the right-hand side above is zero. The quadratic variation of the log-price process over period $t+1$ is defined

$$QV_{t+1} = \operatorname*{plim}_{n\to\infty} \sum_{j=1}^{n} r_{t+j/n}^2, \tag{2}$$

$$r_{t+j/n} = p_{t+j/n} - p_{t+(j-1)/n}$$

where the price series on day $t+1$ is assumed to be observed n times $\left\{p_{t+1/n}, \ldots, p_{t+1-1/n}, p_{t+1}\right\}$. See Andersen et al. (2006), Barndorff-Nielsen and Shephard (2007) and Aït-Sahalia and Jacod (2014) for surveys of volatility estimation and forecasting using high frequency data. The objective of this paper is to compare the variety of estimators of QV that have been proposed in the literature to date. We do so with emphasis on comparisons with the simple realized variance estimator, which is the empirical analog of QV:

$$RV_{t+1} = \sum_{j=1}^{n} r_{t+j/n}^2. \tag{3}$$

2.1. Sampling frequency, sampling scheme, and sub-sampling

We consider a variety of classes of estimators of asset price variability. All realized measures require a choice of sampling frequency (e.g., 1-second or 5-minute sampling), sampling scheme (calendar time or tick time), whether to use transaction prices or mid-quotes, when both are available. Thus even for a very simple estimator such as realized variance, there are a number of choices to be made. To examine the sensitivity of realized measures to these choices, we implement each measure using calendar-time sampling of 1 s, 5 s, 1 min, 5 min and 15 min. We also consider tick-time sampling using samples that yield *average* durations that match the values for calendar-time sampling, as well as a "tick-by-tick" estimator that simply uses every available observation. Subsampling,[1] introduced by Zhang et al. (2005), is a simple way to improve efficiency of some sparse-sampled estimators. We consider subsampled versions of all estimators (except estimators using tick-by-tick data, which cannot be subsampled).[2] The subsampled version of RV (which turns out to perform very well in our analysis) was first studied as the "second best" estimator in Zhang et al. (2005).

In total we have 5 calendar-time implementations, 6 tick-time implementations, and $5+6-1=10$ corresponding subsampled implementations, yielding 21 realized measures for a given price series. Estimating these on both transaction and quote prices

[1] Subsampling involves using multiple "grids" of prices sampled at a given frequency to obtain a collection of realized measures, which are then averaged to yield the "subsampled" version of the estimator. For example, 5-minute RV can be computed using prices sampled at 9:30, 9:35, etc. and can also be computed using prices sampled at 9:31, 9:36, etc.

[2] In general, we implement subsampling using a maximum of 10 partitions.

yields a total of 42 versions of each realized measure. Of course, some of these combinations are expected to perform poorly empirically (given the extant literature on microstructure biases and the design of some of the estimators described below), and by including them in our analysis we thus have an "insanity check" on whether our tests can identify these poor estimators.

2.2. Classes of realized measures

The first class of estimators is standard realized variance (RV), which is the sum of squared intra-daily returns. This simple estimator is the sample analog of quadratic variation, and in the absence of noisy data, it is the nonparametric maximum likelihood estimator, and so is efficient, see Andersen et al. (2001b) and Barndorff-Nielsen (2002). However, market microstructure noise induces serial auto-correlation in the observed returns, which biases the realized variance estimate at high sampling frequencies (see Hansen and Lunde (2006b) for a detailed analysis of the effects of microstructure noise). When RV is implemented in practice, the price process is often sampled sparsely to strike a balance between increased accuracy from using higher frequency data and the adverse effects of microstructure noise. Popular choices include 1-minute, 5-minute (as in the title of this paper), or 30-minute sampling.

We next draw on the work of Bandi and Russell (2008), who propose a method for optimally choosing the sampling frequency to use with a standard RV estimator. This sampling frequency is calculated using estimates of integrated quarticity[3] and variance of the microstructure noise. These authors also propose a bias-corrected estimator that removes the estimated impact of market microstructure noise. Since the key characteristic of the Bandi–Russell estimator is the estimated optimal sampling frequency, we do not vary the sampling frequency when implementing it. This reduces the number of versions of this estimator from 42 to 8.[4]

The third class of realized measures we consider is the first-order autocorrelation-adjusted RV estimator (RVac1) used by French et al. (1987) and Zhou (1996), and studied extensively by Hansen and Lunde (2006b). This estimator was designed to capture the effect of autocorrelation in high frequency returns induced by market microstructure noise.

The fourth class of realized measures includes the two-scale realized variance (TSRV) of Zhang et al. (2005) and the multi-scale realized variance (MSRV) of Zhang (2006). These estimators compute a subsampled RV on one or more slower time scales (lower frequencies) and then combine with RV calculated on a faster time scale (higher frequency) to correct for microstructure noise. Under certain conditions on the market microstructure noise, these estimators are consistent at the optimal rate. In our analysis, we set the faster time scale by using one of the 21 sampling frequency/sampling scheme combinations mentioned above, while the slower time scale(s) are chosen to minimize the asymptotic variance of the estimator using the methods developed in the original papers. It is worth noting here that "subsampled RV", which we have listed in our first class of estimators, corresponds to the "second-best" form of TSRV in Zhang et al. (2005), in that it exploits the gains from subsampling but does not attempt to estimate and remove any bias in this measure. We keep any measure involving two or more time scales in the TSRV/MSRV class, and any measure based on a single time scale is listed in the RV class.

The fifth class of realized measures is the realized kernel (RK) estimator of Barndorff-Nielsen et al. (2008). This measure is a generalization of RVac1, accommodating a wider variety of microstructure effects and leading to a consistent estimator. Barndorff-Nielsen et al. (2008) present realized measures using several different kernels, and we consider RK with the "flat top" versions of the Bartlett, cubic, and modified Tukey–Hanning$_2$ kernel, and the "non-flat-top" Parzen kernel. The Bartlett and cubic RK estimators are asymptotically equivalent to TSRV and MSRV, respectively, and modified Tukey–Hanning$_2$ was the recommended kernel in Barndorff-Nielsen et al. (2008) in their empirical application to GE stock returns. The non-flat-top Parzen kernel was studied further in Barndorff-Nielsen et al. (2011) and results in a QV-estimator that is always positive while allowing for dependence and endogeneity in the microstructure noise. We implement these realized kernel estimators using the 21 sampling frequency/sampling scheme combinations mentioned above, and estimate the optimal bandwidths for these kernels separately for each day, using the methods in Barndorff-Nielsen et al. (2011). The realized kernel estimators are not subsampled because Barndorff-Nielsen et al. (2011) report that for "kinked" kernels such as the Bartlett kernel, the effects of subsampling are neutral, while for the other three "smooth" kernels, subsampling is detrimental. (The RVac1 measure corresponds to the use of a "truncated" kernel, and subsampling improves performance, so we include the subsampled versions of RVac1 in the study.)

The sixth class of estimators are pre-averaged realized variances (RVpa) estimators first introduced by Podolskij and Vetter (2009a) and further studied in Jacod et al. (2009). Pre-averaging applies a kernel-like weighting function to observed returns to construct pre-averaged returns. The most common weighting function has a Bartlett kernel-like tent shape and so it is identical to first locally averaging prices and then constructing returns as the difference between two adjacent pre-averaged prices. Pre-averaged realized variance and realized kernels are closely related, and in general only differ in the treatment of edge effects—the handling of the first and last few observations. Given a particular sampling scheme, RVpa is implemented using all (overlapping) pre-averaged returns. Following the empirical work of Christensen et al. (2014) and Hautsch and Podolskij (2013), we use a kernel bandwidth $K = \lceil \theta\sqrt{n} \rceil$, where $\theta = 1$ and n is the number of sampled intraday returns.

The seventh class of estimators is the "realized range-based variance" (RRV) of Christensen and Podolskij (2007) and Martens and Van Dijk (2007). Early research by Parkinson (1980), Andersen and Bollerslev (1998) and Alizadeh et al. (2002) show that the properly scaled, daily high-low range of log prices is an unbiased estimator of daily volatility when constant, and is more efficient than squared daily open-to-close returns. Correspondingly, Christensen and Podolskij (2007) and Martens and Van Dijk (2007) apply the same arguments to intraday data, and improve on the RV estimator by replacing each intraday squared return with the high-low range from a block of intra-day returns. To implement RRV, we use the sampling schemes described above, and then use block size of 5, following Patton and Sheppard (2009b), and block size of 10, which is close to the average block size used in Christensen and Podolskij's application to General Motors stock returns.

Finally, we include the maximum likelihood Realized Variance (MLRV) of Aït-Sahalia et al. (2005), which assumes that the observed price process is composed of the efficient price plus i.i.d. noise such that the observed return process follows an MA(1) process with parameters that can be estimated using Gaussian

[3] Estimates of daily integrated quarticity are estimated using 39 intra-day prices sampled uniformly in tick-time.

[4] Note that the Bandi–Russell RV measure is not consistent for QV in the presence of jumps. (This is also the case for realized range and MLRV, described below.) We include these estimators in our comparison as these are widely-used and cited realized measures, and we leave it to the data to shed light on which estimators perform well empirically.

MLE. This estimator is shown to be robust to misspecification of the marginal distribution of the microstructure noise by Aït-Sahalia et al. (2005), but is sensitive to the independence assumption of noise, as demonstrated in Gatheral and Oomen (2010).

The total number of realized measures we compute for a single price series is 210, so an asset with both transactions and quote data has a set of 420 realized measures.[5,6]

2.3. Additional realized measures

Our main empirical analysis focuses on realized measures that estimate the quadratic variation of an asset price process. From a forecasting perspective, work by Andersen et al. (2007) and others has shown that there may be gains to decomposing QV into the component due to continuous variation (integrated variance, or IV) and the component due to jumps (denoted JV):

$$QV_{t+1} = \operatorname*{plim}_{n\to\infty} \sum_{j=1}^{n} r^2_{t+j/n} = \underbrace{\int_t^{t+1} \sigma^2(s)\, ds}_{IV_{t+1}} + \underbrace{\sum_{t<s\le t+1} \kappa^2(s)}_{JV_{t+1}}. \quad (4)$$

Thus for our forecasting application in Section 5.6, we also consider four classes of realized measures that are "jump robust", i.e., they estimate IV not QV. The first of these is the bi-power variation (BPV) of Barndorff-Nielsen and Shephard (2006), which is a scaled sum of products of adjacent absolute returns. We also estimate a pre-averaged version of Bipower Variation, motivated by Podolskij and Vetter (2009b) and Christensen et al. (2014).

The second class of jump-robust realized measures is the quantile-based realized variance (QRV) of Christensen et al. (2010). QRV is based on combinations of locally extreme quantile observations within blocks of intra-day returns, and requires choice of block length and quantiles. It is reported to have better finite sample performance than BPV in the presence of jumps, and additionally is consistent, efficient and jump-robust even in the presence of microstructure noise. For implementation, we use the asymmetric version of QRV with rolling overlapping blocks[7] and quantiles approximately equal to 0.85, 0.90 and 0.96, following the authors' application to Apple stock returns. The block lengths are chosen to be around 100, with the exact value depending on the number of filtered daily returns, and the quantile weights are calculated optimally following the method in Christensen et al. (2010). QRV is the most time-consuming realized measure to estimate, and thus is not further subsampled.

The third class of jump-robust realized measures are the "nearest neighbor truncation" estimators of Andersen et al. (2012), specifically their "MinRV" and "MedRV" estimators. These are the scaled square of the minimum of two consecutive intra-day absolute returns or the median of 3 consecutive intra-day absolute returns. These estimators are more robust to jumps and microstructure noise than BPV, and MedRV is designed to handle outliers or incorrectly entered price data.

The final class of jump-robust measures estimators is the truncated or threshold realized variance (TRV) of Mancini (2001, 2009), which is the sum of squared returns, but only including returns that are smaller in magnitude than a certain threshold. Following Corsi et al. (2010), we take the threshold to be three times a local (intra-day) volatility estimate.[8]

In total, across sampling frequencies and subsampling/not subsampling we include 228 jump-robust realized measures in our forecasting application, in addition to the 420 estimators described in the previous section.

3. Comparing the accuracy of realized measures

We examine the empirical accuracy of our set of competing measures of asset price variability using two complementary approaches.

3.1. Comparing estimation accuracy

We first compare the accuracy of realized measures in terms of their estimation error for a given day's quadratic variation. QV is not observable, even *ex post*, and so we cannot directly calculate a metric like mean-squared error and use that for the comparison. We overcome this by using the data-based ranking method of Patton (2011a). This approach requires employing a proxy (denoted $\tilde{\theta}$) for the quadratic variation that is assumed to be unbiased, but may be noisy.[9] This means that we must choose a realized measure that is unlikely to be affected by market microstructure noise. Using proxies that are more noisy will reduce the ability to discriminate between estimators, but will not affect consistency of the procedure. We use the squared open-to-close returns from transaction prices (RVdaily) for our main analysis, and further consider 15-minute RV, 5-minute RV, 1-minute MSRV and 1-minute RKth2, all computed on transaction prices using tick-time sampling, as possible alternatives.[10] Since estimators based on the same price data are correlated, it is necessary to use an instrument for the proxy to "break" the dependence between the estimation error in the realized measure under analysis and the estimation error in the proxy. As our instrument we use a one-day lead of the proxy.[11]

[5] Specifically, for each of RV, RVpa, TSRV, MSRV, MLRV, RVac1, RRV (with two choices of block size) and RK (with 4 different kernels), 11 not-subsampled estimators, which span different sampling frequencies and sampling schemes, are implemented on each of the transactions and midquotes price series. In addition, we estimate 2 bias-corrected Bandi–Russell realized measures and 2 not-bias-corrected BR measures (calendar-time and tick-time sampling) per price series. These estimators account for $12\times11\times2+(2+2)\times2 = 272$ of the total set. RV, TSRV, MSRV, MLRV, RVac1 and RRV ($m = 5$ and 10) also have 10 subsampled estimators per price series, and there are 4 subsampled BR estimators per price series, which adds $7\times10\times2+4\times2 = 148$ subsampled estimators to the set. In total, this makes $272 + 148 = 420$ estimators.

[6] Research on estimating volatility using high-frequency data has continued since this project began, and some new estimators have recently been proposed that are not included in our analysis, e.g., the estimators of Bibinger et al. (2014) and Jacod and Todorov (2014).

[7] Christensen et al. (2010) refers to this formulation of the QRV as "subsampled QRV", as opposed to "block QRV", which has adjacent non-overlapping blocks. However, we do not use this terminology as this type of "subsampling" is different from the subsampling we implement for the other estimators.

[8] The algorithm to compute local volatility can fail when very high-frequency sampling is applied to days with low liquidity, and in such cases we consider the TRV to be non-implementable. These occurrences are generally limited to 1-second sampling or, less often, 5-second sampling of illiquid assets such as computed indices or individual equities.

[9] Numerous estimators of quadratic variation can be shown to be asymptotically unbiased as the sampling interval goes to zero, however this approach requires unbiasedness for a fixed sampling interval.

[10] These four additional proxies were found to be unbiased for the RVdaily measure for the majority of assets, and in addition, are generally much more precise.

[11] As described in Patton (2011a), the use of a lead (or lag) of the proxy formally relies on the daily quadratic variation following a random walk. Numerous papers, see Bollerslev et al. (1994) and Andersen et al. (2006) for example, find that conditional variance is a highly persistent process, close to being a random walk. Hansen and Lunde (2014) study the quadratic variation of all 30 constituents of the Dow Jones Industrial Average and reject the null of a unit root for few of the stocks. Meddahi (2003) shows analytically that certain classes of continuous time stochastic volatility processes imply that their daily integrated variance follows an ARMA process, with autoregressive persistence governed by the persistence of the spot variance. Simulation results in Patton (2011a) show that inference based on a random walk approximation has acceptable finite-sample properties for DGPs that

The comparison of estimation accuracy also, of course, requires a metric for measuring accuracy. The approach of Patton (2011a) allows for a variety of metrics, including the MSE and QLIKE loss functions. Simulation results in Patton and Sheppard (2009a), and empirical results in Hansen and Lunde (2005), Patton and Sheppard (2009b) and Patton (2011a) all suggest that using QLIKE leads to more power to reject inferior estimators.[12] The QLIKE loss function is defined as:

$$\text{QLIKE } L(\theta, M) = \frac{\theta}{M} - \log\frac{\theta}{M} - 1 \tag{5}$$

where θ is QV, or a proxy for it, and M is a realized measure. With this in hand, we obtain a consistent (as $T \to \infty$) estimate of the difference in accuracy between any two realized measures:

$$\frac{1}{T}\sum_{t=1}^{T} \Delta\tilde{L}_{ij,t} \xrightarrow{p} E\left[\Delta L_{ij,t}\right] \tag{6}$$

where $\Delta\tilde{L}_{ij,t} \equiv L\left(\tilde{\theta}_t, M_{it}\right) - L\left(\tilde{\theta}_t, M_{jt}\right)$ and $\Delta L_{ij,t} \equiv L(\theta_t, M_{it}) - L\left(\theta_t, M_{jt}\right)$. Under standard regularity conditions (see Patton (2011a) for example) we can use a block bootstrap to conduct tests on the estimated differences in accuracy, such as the pair-wise comparisons of Diebold and Mariano (2002) and Giacomini and White (2006), the "reality check" of White (2000) as well as the multiple testing procedure of Romano and Wolf (2005), and the "model confidence set" of Hansen et al. (2011).

3.2. Comparing forecast accuracy

The second approach we consider for comparing realized measures is through a simple forecasting model. As we describe in Section 5.6, we construct volatility forecasts based on the heterogeneous autoregressive (HAR) model of Corsi (2009), estimated separately for each realized measure. The problem of evaluating volatility forecasts has been studied extensively, see Hansen and Lunde (2005, 2006a), Andersen et al. (2005), and Patton (2011b) among several others. The latter two papers focus on applications where an unbiased volatility proxy is available, and again under standard regularity conditions we can use block bootstrap methods to conduct tests such as those of Diebold and Mariano (2002), White (2000), Romano and Wolf (2005), Giacomini and White (2006), and Hansen et al. (2011).

4. Data description

We use high frequency (intra-daily) asset price data for 31 assets spanning five asset classes: individual equities (from the US and the UK), equity index futures, computed stock indices, currency futures and interest rate futures. The data are transactions and quotations prices taken from Thomson Reuter's Tick History. The sample period is January 2000 to December 2010, though data availability limits us to a shorter sub-period for some assets. Short days, defined as days with prices recorded for less than 60% of the regular market operation hours, are omitted. For each asset, the number of short days is small compared to the total number of days—the largest proportion of days omitted is 1.7% for ES (E-mini S&P500 futures). Across assets, we have an average of 2537 trading days, with the shortest sample being 1759 trade days (around 7 years) and the longest 2782 trade days. All series were cleaned according to a set of baseline rules similar to those in Barndorff-Nielsen et al. (2009). Data cleaning details are provided in the appendix.[13]

Table 1 presents the list of assets, along with their sample periods and some summary statistics. Computed stock indices are not traded assets and are constructed using trade prices, and so quotes are unavailable. This table reveals that these assets span not only a range of asset classes, but also characteristics: average annualized volatility ranges from under 2%, for interest rate futures, to over 40%, for individual equities. The average time between price observations ranges from under one second, for the E-mini S&P 500 index futures contract, to nearly one minute, for some individual equities and computed equity indices.[14]

Given the large number of realized measures and assets, it is not feasible to present summary statistics for all possible combinations. Table A1 in the appendix describes the shorthand used to describe the various estimators,[15] and in Table 2 we present summary statistics for a selection of realized measures for two assets, Microsoft and the US dollar/Australian dollar futures contract.[16] Tables A3 and A4 in the appendix contain more detailed summary statistics. Table 2 reveals some familiar features of realized measures: those based on daily squared returns have similar averages to realized measures using high (but not too high) frequency data, but are more variable, reflecting greater measurement error. For Microsoft, for example, RVdaily has an average of 3.20 (28.4% annualized) compared with 3.37% for RV5min, but its standard deviation is more than 25% larger than that of RV5min. We also note that RV computed using tick-by-tick sampling (i.e., the highest possible sampling) is much larger on average than the other estimators, more than 3 times larger for Microsoft and around 50% larger for the USD/AUD exchange rate, consistent with the presence of market microstructure noise. This distortion vanishes when pre-averaging is used.

In the last four columns of Table 2 we report the first- and second-order sample autocorrelations of the realized measures, as well as estimates of the first- and second-order autocorrelation of the underlying quadratic variation using the estimation method in Hansen and Lunde (2014).[17] As expected, the latter estimates are much higher than the former, reflecting the attenuation bias due to the estimation error in a realized measure. Using the method of Hansen and Lunde (2014), the estimated first-order autocorrelation of QV for Microsoft and the USD/AUD exchange rate is around 0.95, while the sample autocorrelations for the realized measures themselves average around 0.68. Table A4

are persistent but strictly not random walks, and we confirm in Table A4, in the appendix, that all series studied here are highly persistent. In Section 5 we also present results based on an AR(1) approximation rather than a random walk. We also consider the use of a one-day lag of the proxy, and find the results (reported in the appendix) to be very similar to our base case using a one-day lead.

12 We also present some results based on the MSE loss function. See Section 5 and the online appendix.

13 The sensitivity of estimators in different classes to data cleaning methods is an interesting topic, as is developing estimators that are more robust to various data cleaning rules. We do not explore these issues. We note here that the data provided by Thomson Reuter's Tick History, especially the data on futures, is very clean compared with the more-widely used NYSE TAQ data.

14 Most futures contracts trade nearly 24 h a day. However, their liquidity is typically concentrated around a relatively short interval, usually less than half of the day. We measured the percentage of trades that occurred in five minute block of the day using local time-stamps to avoid issues with daylight saving changes, and selected the largest contiguous block where the percentage of observations in the block was above 20% of 1/288.

15 For example, "RV_1m_ct_ss" refers to realized variance (RV), computed on 1-minute data (1m) sampled in calendar time (c), using trade prices (t), with subsampling (ss). See Table A1 for details.

16 All realized measures were computed using code based on Kevin Sheppard's "Oxford Realized" toolbox for Matlab, https://www.kevinsheppard.com/MFEToolbox.

17 Following their empirical application to the 30 DJIA stocks, we use the demeaned 4th through 10th lags of the daily QV estimator as instruments.

Table 1
Description of assets and price series.

Assets		Dates	T	Avg. Ann. Vol.	Avg. Trade Dur.	Avg. Quote Dur.
US equities (NYSE)						
KO	Coca Cola	1/3/2000–12/31/2010	2766	18.8	7.6	2.6
LSI	LSI corp.	1/3/2000–12/31/2010	2767	48.5	15.6	3.8
MSFT	Microsoft	1/3/2000–12/31/2010	2763	24.5	2.7	1.5
IFF	Intl. Flavors & Fragrances	1/3/2000–12/31/2010	2767	23.9	26.6	5.4
SYY	Sysco	1/3/2000–12/31/2010	2766	22.1	12.5	3.4
UK equities (LSE)						
DGE	Diageo	1/4/2000–12/31/2010	2769	23.9	15.8	3.6
VOD	Vodaphone	1/4/2000–12/31/2010	2770	29.5	7.0	2.3
SAB	SABMiller	1/4/2000–12/31/2010	2733	27.9	23.6	3.8
SDR	Schroders	1/4/2000–12/31/2010	2757	45.8	52.4	8.7
RSA	RSA Ins.	1/4/2000–12/31/2010	2768	39.1	28.1	6.4
Interest rate futures						
TU	2 yr Treasury note	1/2/2003–12/31/2010	1994	1.4	7.6	0.5
FV	5 yr Treasury note	1/2/2001–12/31/2010	2486	3.5	3.0	0.3
TY	10 yr Treasury note	1/2/2001–12/31/2010	2484	5.2	1.9	0.3
US	30 yr Treasury bond	1/2/2001–10/29/2010	2449	8.1	2.4	0.4
FGBS	German short term govt bond	1/3/2000–10/29/2010	2735	1.3	9.0	1.9
FGBL	German long term govt bond	1/3/2000–10/29/2010	2741	4.6	2.7	1.0
Currency futures						
CD	Canadian Dollar	1/2/2004–12/31/2010	1763	8.4	4.1	0.6
AD	Australian Dollar	1/2/2004–12/30/2010	1759	9.3	4.9	0.5
BP	British Pound	1/2/2004–12/31/2010	1762	6.7	2.9	0.4
URO	Euro	1/2/2004–12/31/2010	1762	6.9	1.4	0.3
JY	Japanese Yen	1/2/2004–12/31/2010	1763	7.3	3.1	0.4
Index futures						
STXE	EuroStoxx50	1/3/2000–12/30/2010	2782	17.9	2.0	0.7
JNI	Nikkei 225	1/4/2000–10/29/2010	2644	15.2	3.5	0.9
FDX	DAX 40	1/3/2000–10/29/2010	2738	17.9	1.5	0.8
FFI	FTSE 100	1/4/2000–10/29/2010	2707	15.6	1.9	0.5
ES	e-mini S&P 500	1/3/2000–12/31/2010	2750	14.6	0.5	0.2
Market indices						
SPX	S&P500	1/3/2000–12/31/2010	2719	16.1	15.9	–
STOXX50E	EuroStoxx50	1/3/2000–12/30/2010	2782	18.6	15.2	–
DAX	DAX 40	1/4/2006–12/30/2010	2781	19.4	2.9	–
FTSE	FTSE 100	1/4/2000–12/31/2010	2762	15.9	4.9	–
N225	Nikkei 225	1/5/2000–12/30/2010	2665	14.7	48.1	–

Notes: This table presents the 31 assets included in the analysis, the sample period and number of trading days for each asset, and some summary statistics: the average volatility (annualized, estimated using squared open-to-close returns), and the average trade and quote durations (in seconds).

Table 2
Summary statistics of some sample realized measures for two representative assets.

	Mean	Std. dev.	Skew	Kurt	Min	Max	$ac(1)$	$ac(2)$	$ac^*(1)$	$ac^*(2)$
Microsoft (MSFT)										
RVdaily	3.20	7.21	6.53	72.09	0.00	112.86	0.26	0.29	0.96	0.99
RV_5m_ct	3.37	4.48	4.56	36.86	0.18	63.14	0.72	0.68	0.96	0.95
RV_5m_ct_ss	3.27	4.38	4.84	44.17	0.17	71.69	0.72	0.68	0.96	0.95
RV_1t_bt	11.24	20.36	3.75	20.96	0.27	207.58	0.94	0.92	0.99	0.98
RVpa_1t_bt	3.27	4.28	5.78	71.95	0.19	86.52	0.72	0.72	0.94	0.92
RVac1_1m_ct	3.40	4.54	5.22	53.70	0.15	81.89	0.72	0.70	0.94	0.94
RKth2_1m_bt	3.32	4.47	4.71	40.64	0.09	69.96	0.71	0.67	0.95	0.95
MSRV_1m_ct	3.23	4.51	4.81	41.16	0.13	68.19	0.69	0.65	0.96	0.95
MLRV_5s_ct	3.21	3.62	5.02	50.41	0.26	63.32	0.80	0.77	0.95	0.93
RRVm5_1m_ct	3.34	4.23	5.37	61.72	0.21	81.49	0.74	0.72	0.94	0.93
USD/AUD exchange rate future (AD)										
RVdaily	0.46	1.37	9.88	149.55	0.00	28.95	0.39	0.40	0.98	0.93
RV_5m_ct	0.52	1.05	7.90	91.46	0.04	17.21	0.71	0.78	0.94	0.93
RV_5m_ct_ss	0.51	1.02	7.69	86.66	0.04	15.77	0.74	0.80	0.92	0.91
RV_1t_bt	0.70	1.04	7.61	92.73	0.07	18.37	0.70	0.70	0.95	0.91
RVpa_1t_bt	0.51	1.01	7.81	91.26	0.04	16.70	0.76	0.79	0.94	0.91
RVac1_1m_ct	0.52	1.02	7.95	96.27	0.04	18.14	0.73	0.78	0.94	0.93
RKth2_1m_bt	0.51	1.04	8.44	107.35	0.04	17.70	0.71	0.78	0.92	0.90
MSRV_1m_ct	0.51	1.04	8.06	95.30	0.04	17.04	0.72	0.79	0.92	0.91
MLRV_5s_ct	0.57	0.99	6.91	71.92	0.06	16.06	0.79	0.78	0.96	0.92
RRVm5_1m_ct	0.54	1.00	7.29	78.92	0.05	16.25	0.78	0.79	0.95	0.91

Notes: This table displays the summary statistics for several estimators for Microsoft an Australian–US Dollar futures. Referring to the four right-most columns, $ac(p)$ denotes the pth sample autocorrelation, and $ac^*(p)$ denotes the pth estimated autocorrelation of QV based on a realized measure, using the instrumental variables method of Hansen and Lunde (2014).

presents summaries of these autocorrelations for all 31 assets, and reveals that the estimated first- (second-) order autocorrelation of the underlying QV is high for all assets. The average estimate across all assets and realized measures, even including poor estimators, equals 0.95 (0.92). These findings support our use, in the next section, of the ranking method of Patton (2011a), which relies on high persistence of QV.

5. Empirical results on the accuracy of realized measures

We now present the main analysis of this paper. We firstly discuss simple rankings of the realized measures, and then move on to more sophisticated tests to formally compare the various measures. As described in Section 3, we measure accuracy using the QLIKE distance measure, using squared open-to-close returns (RVdaily) as the volatility proxy, with a one-day lead to break the dependence between estimation error in the realized measure and error in the proxy. In some of the analysis below we consider using higher frequency RV measures for the proxy (RV15min and RV5min), as well as some non-RV proxies, namely 1-minute MSRV and 1-minute Tukey–Hanning$_2$ realized kernel.

5.1. Rankings of average accuracy

We firstly present a summary of the rankings of the 420 realized measures applied to the 31 assets in our sample. These rankings are based on average, unconditional distance of the measure from the true QV, and in Section 5.5 we consider conditional rankings.

The top panel of Table 3 presents the "top 10" individual realized measures, according to their average rank across all assets in a given class.[18] It is noteworthy that 5-minute RV does *not* appear in the top 10 for any of these asset classes. This is some initial evidence that there are indeed better estimators of QV available, and we test whether this outperformance is statistically significant in the sections below.

With the caveat that these estimated rankings do not come with any measures of significance, and that realized measures in the same class are likely highly correlated,[19] we note the following patterns in the results. Realized kernels appear to do well for individual equities (taking 4 and 5 of the top 10 slots, respectively, for US and UK equities), realized range does well for interest rate futures (8 out of top 10), and two/multi-scales RV do well for currency futures (6 out of the top 10). For computed indices, RVac1 and realized kernels comprise the entire top 10. The top 10 realized measures for index futures contain a smattering of measures across almost all classes. The lower panel of Table 3 presents a summary of the upper panel, sorting realized measures by class and sampling frequency.

It is perhaps also interesting to note which price series is most often selected. We observe a mix of trades and quotes for individual equities,[20] while we see mid-quotes dominating the top 10 for interest rate futures and currency futures. For equity index futures, transaction prices make up the entire top 10. (Our computed indices are only available with transaction prices, so no comparisons are available for that asset class.)

5.2. Pair-wise comparisons of realized measures

To better understand the characteristics of a "good" realized measure, we present results on pair-wise comparisons of measures that differ only in one aspect. We consider four features: the use of transaction prices vs. mid-quotes; the use of calendar-time vs. tick-time sampling; the use of subsampling; and the use of pre-averaging. For each class of realized measures and for each sampling frequency, we compare pairs of estimators that differ in one of these dimensions, and compute a robust t-statistic on the average difference in loss, separately for each asset.[21] Table 4 presents the proportion (across the 31 assets) of t-statistics that are significantly positive minus the proportion that are significantly negative.[22] A negative entry in a given element indicates that the first approach (e.g., transaction prices, in the top panel) outperforms the second approach.

The top panel of Table 4 shows that for these assets, transaction prices are generally preferred to quote prices for most estimator-frequency combinations, especially at lower frequencies. This is unsurprising since at these frequencies the sampling frequency limits the effects of bid–ask bounce microstructure noise. Exceptions are RV, MLRV and RRV at the highest frequencies (1-tick and/or 1-second) and MSRV at low frequencies. Most noise robust estimators prefer transaction prices, which is consistent with these estimators being designed specifically to mitigate the effect of transaction price noise.

The second panel of Table 4 reveals that for high frequencies (1-second and 5-second), calendar time sampling is preferred to tick-time sampling, while for lower frequencies (5-minute and 15-minute), tick-time sampling generally leads to better realized measures. Oomen (2006b) and Hansen and Lunde (2006b) provide theoretical grounds for why tick-time sampling should outperform calendar-time sampling, and at lower frequencies this appears to be true. Microstructure noise, and in particular the dependence in the noise, likely plays a role at the highest frequencies, and the ranking of calendar-time and tick-time sampling depends on their sensitivity to this noise. RVpa and RK appear to be insensitive to the sampling scheme at the highest frequencies.

The third panel of Table 4 compares realized measures with and without subsampling. Theoretical work by Zhang et al. (2005) and Zhang (2006) suggests that subsampling is a simple way to improve the efficiency of a realized measure. Our empirical results generally confirm that subsampling is helpful, at least when using lower frequency (5-minute and 15-minute) data. For higher frequencies (1-second to 1-minute), subsampling has either no effect or a negative effect on accuracy. Interestingly, we note that for realized range (RRV), subsampling reduces accuracy across all sampling frequencies.

Finally, the bottom panel presents clear advice on pre-averaging: it is beneficial at the highest sampling frequencies (5 s or less) and harmful at frequencies lower than one minute. This is consistent with the theoretical underpinnings of pre-averaging, see Podolskij and Vetter (2009a) and Jacod et al. (2009), both of which suggest applying pre-averaging to data sampled at the highest frequency.

[18] Table A6 in the appendix presents rank correlation matrices for each asset class, and confirms that the rankings of realized measures for individual assets in a given asset class are relatively consistent, with average within-asset-class rank correlations ranging from 0.70 to 0.87.

[19] See Table A5 in the appendix for a summary of the correlations between realized measures.

[20] In fact, decomposing this group into US equities and UK equities, we see that the top 10 realized measures for US equities all use transaction prices, while the top 10 for UK equities all use mid-quotes, perhaps caused by different forms of market microstructure noise on the NYSE (the exchange for 4 of the 5 US stocks) and the LSE.

[21] This is done as a panel regression for a single asset, as for each measure of a specific estimator class and sampling frequency, there are $2 \times 2 \times 2 = 8$ versions (cal-time vs. tick time, trades vs. quotes, not subsampled vs. subsampled), and conditioning on one of these characteristics leaves 4 versions.

[22] Columns that are not relevant for the comparison have blank values. For example, there is no calendar time equivalent to 1-tick sampling. Additionally, the second panel covers only 26 assets, since there are no quotation prices for the 5 computed indices. Finally, the third panel does not contain the RK row, given the work of Barndorff-Nielsen et al. (2011).

Table 3
Top 10 estimators for each asset class and the average rank within the asset class.

	US equities	avg rank	UK equities	avg rank	Bond futures	avg rank	FX futures	avg rank	Index futures	avg rank	Computed indices	avg rank
	tr RKbart_5s_b	14	mq RKth2_5s_b	6	mq RRVm5_5s_b	29	tr TSRV_1s_c_ss	21	tr RV_1m_b_ss	33	RVac1_1m_b	7
	tr RRVm5_1m_b_ss	14	mq RKbart_5s_b	8	mq RRVm5_5s_b_ss	30	tr TSRV_1s_c	21	tr RVac1_1m_b_ss	33	RVac1_1m_c	8
	tr RVpa_1s_c	15	mq RRVm5_1m_b_ss	9	mq RRVm10_1s_c_ss	34	mq MSRV_1s_b_ss	23	tr MSRV_5s_c_ss	37	RKth2_1t_b	11
	tr RVac1_1m_b_ss	18	mq RVpa_1s_b	10	mq RRVm10_1s_c	35	mq MLRV_1s_c	24	tr RV_1m_b	37	RKcub_1t_b	12
	tr RKth2_5s_c	18	mq RRVm5_1m_b	12	mq RRVm10_1s_b_ss	36	mq MLRV_1s_c_ss	25	tr RKbart_1s_c	37	RKbart_1m_b	14
	tr RRVm5_1m_c_ss	19	mq RKnfp_1s_b	12	mq RRVm5_5s_c_ss	36	mq MSRV_1s_b	25	tr MSRV_5s_c	38	RKbart_1t_b	15
	tr RKbart_5s_c	20	mq RKbart_1s_b	18	mq RRVm5_5s_c	36	mq RV_5s_c	26	tr RKbart_1s_b	41	RKth2_1m_b	15
	tr RRVm5_1m_b	21	mq RV_5m_b_ss	23	mq RRVm10_1s_b	37	mq RV_5s_c_ss	27	tr RKth2_1s_c	41	RKbart_1m_c	16
	tr RRVm5_1m_c	21	mq RKcub_1s_b	24	tr RKth2_1s_c	37	mq MSRV_1s_c	28	tr RVac1_1m_c_ss	41	RKnfp_1t_b	16
	tr RKth2_5s_b	22	mq RVpa_5s_b	31	mq RVac1_5s_c	39	mq MSRV_1s_c_ss	28	tr RRVm10_5s_b_ss	43	RKnfp_5s_b	17
Asset class	Freq.	No.	Freq.	No.	Freq.	No.	Freq.	No.	Freq.	No.	Freq.	No.
RV		0	5 m	1		0	5 s	2	1 m	2		0
RVac1	1 m	1		0	5 s	1		0	1 m	2	1 m	2
RK												
bart	5 s	2	1 s, 5 s	2		0		0	1 s	2	1t,1 m	3
cubic		0	1 s	1		0		0		0	1t	1
th2	5 s	2	5 s	1	1 s	1		0	1 s	1	1t,1 m	2
nfp		0	1 s	1		0		0		0	1t,5 s	2
kSRV												
tsrv		0		0		0	1 s	2		0		0
msrv		0		0		0	1 s	4	5 s	2		0
MLRV		0		0		0	1 s	2		0		0
RRV												
$\boldsymbol{m} = 5$	1 m	4	5 m	2	5 s	4		0		0		0
$\boldsymbol{m} = 10$		0		0	1 s	4		0	5 s	1		0
RVpa	1 s	1	1 s, 5 s	2		0		0		0		0

Notes: For each asset class, we take the average of the rankings from all assets of that class. The top panel of this table lists the estimators with the top "average-ranks" for each asset class. The bottom panel summarizes the top panel by categorizing them by estimator characteristics.

5.3. Does anything beat 5-minute RV?

Realized variance, computed with a reasonable choice of sampling frequency, is often taken as a benchmark or rule-of-thumb estimator for volatility, see Andersen et al. (2001a) and Barndorff-Nielsen (2002) for example. This measure has been used as far back as French et al. (1987), is simple to compute, and when implemented on a relatively low sampling frequency (such as five minutes) requires much less data and cleaning.[23] Thus it is of great interest to know whether it is significantly outperformed by one of the many more sophisticated realized measures proposed in the literature.

We use the stepwise multiple testing method of Romano and Wolf (2005) to address this question. The Romano–Wolf method tests the unconditional accuracy of a set of estimators relative to that of a benchmark realized measure, which we take to be RV computed using 5-minute calendar-time sampling on transaction prices (which we denote RV5min). This procedure is an extension of the "reality check" of White (2000), allowing us to determine not only *whether* the benchmark measure is rejected, but to identify the competing measures that led to the rejection. Formally, the Romano–Wolf stepwise method examines the set of null hypotheses:

$$H_0^{(s)} : E\left[L\left(\theta_t, M_{t,0}\right)\right] = E\left[L\left(\theta_t, M_{t,s}\right)\right], \quad \text{for } s = 1, 2, \ldots, S \quad (7)$$

and looks for realized measures, $M_{t,s}$, such that either $E\left[L\left(\theta_t, M_{t,0}\right)\right] > E\left[L\left(\theta_t, M_{t,s}\right)\right]$ or $E\left[L\left(\theta_t, M_{t,0}\right)\right] < E\left[L\left(\theta_t, M_{t,s}\right)\right]$. The Romano–Wolf procedure controls the "family-wise error rate", which is the probability of making one or more false rejections among the set of hypotheses. We run the Romano–Wolf test in both directions, first to identify the set of realized measures that are significantly worse than RV5min, and then to identify the set of realized measures that are significantly better than RV5min. We implement the Romano–Wolf procedure using the Politis and Romano (1994) stationary bootstrap with 1000 bootstrap replications and an average block size of 10 days.[24] A summary of results is presented in Table 5, and detailed results can be found in the online appendix.

The striking feature of Table 5 is the preponderance of estimators that are significantly beaten by RV5min, and the almost complete lack of estimators that significantly beat RV5min. Concerns about potential low power of this inference method are partially addressed by the ability of this method to reject so many estimators as significantly worse than RV5min: using RVdaily as the proxy we reject an average of 193 estimators (out of 420) as significantly worse than RV5min, which represents almost half of the set of competing measures.[25]

Table 4
Pairwise comparison of realized measures.

Transaction prices vs Mid-quote prices

	1t	1s	5s	1m	5m	15m
RV	69	73	69	-23	-54	-50
RVss		73	69	-15	-65	-96
RVpa	19	-42	-65	-88	-88	-85
RVac1	0	77	50	-23	-38	-4
RK	13	-35	-47	-52	-35	-13
M/TSRV	21	8	-31	-67	-31	-10
MLRV	-31	77	23	-50	-46	4
RRV	8	27	-15	-65	-96	-85
BR	23					

Calendar-time sampling vs Tick-time sampling

	1t	1s	5s	1m	5m	15m
RV		-84	-74	3	23	29
RVss		-84	-74	0	32	23
RVpa		-3	35	65	74	71
RVac1		-84	-65	-3	45	29
RK		-9	-1	48	58	36
M/TSRV		-40	-35	15	37	40
TSRV		-35	-52	0	32	45
MLRV		-81	-45	6	42	19
RRV		-61	3	58	81	77
BR		3				

Not Subsampled vs Subsampled estimators

	1t	1s	5s	1m	5m	15m
RV		5	7	6	29	48
RVac1		-86	-46	29	84	94
M/TSRV		-2	0	13	40	19
MLRV		0	4	10	68	77
RRV		-13	-32	-19	-45	-58
BR		6				

Not Pre-averaged RV vs Pre-averaged RV

	1t	1s	5s	1m	5m	15m
RV	74	77	39	-74	-100	-100

Notes: This table summarizes results on pairwise comparisons of realized measures that differ only in the price series used (top panel), sampling scheme used (middle panel), or use of subsampling or pre-averaging (bottom two panels). For each pair, a robust t-statistic on the average loss difference is computed per asset and estimator type. Each table cell summarizes the pairwise comparisons for a given estimator class and frequency by reporting the proportion of significantly positive t-statistics minus the proportion of significantly negative t-statistics. A negative value indicates that the first approach (e.g., calendar-time sampling in the top panel) outperforms the second approach, a positive value indicates the opposite. Values less than −33 are dark-shaded, and values greater than 33 are light-shaded. 'RK' aggregates the resulting t-statistics from the 4 types of Realized Kernels, MSRV and TSRV results are combined in 'M/TSRV', and RRVm5 and RRVm10 results are combined.

We also present results using the other four proxies. These proxies are more precise, although they are potentially more susceptible to market microstructure noise. (Proxies that have an unconditional mean that is significantly different from that of

[23] Of course, a sampling frequency of five minutes is only "relatively low" for liquid assets; for some assets, such as corporate bonds, a five-minute sampling frequency would be quite high. Five-minute sampling has emerged as a rough benchmark in the extant literature since the vast majority of empirical studies look at very liquid assets like exchange rates and US equities. Given that all of our 31 assets are relatively liquid, we adopt five-minute RV as our benchmark estimator.

[24] The ideal choice of block size length is driven by the persistence in the variable we are interested in testing, which in our case is the loss difference, $\Delta L(\tilde{\theta}_t, M_{t,0}, M_{t,s})$. The QLIKE loss, which is based on the realized measure and a proxy, is substantially less persistent than the realized measure, and by taking the difference between two measures and a proxy we further reduce the persistence. To confirm this, we compute the optimal block length, using the method of Politis and White (2004), for all pairs of measures and all assets. The mean optimal block length for loss differences is 4, and the median is 2. In contrast, the mean optimal block length for the measures themselves is 97, and the median is 103. A block length of 10 for the loss differences is thus a reasonably conservative choice.

[25] Note that the Romano-Wolf test controls the family-wise error (FWE) rate, defined as the probability of rejecting a single true null hypothesis across all 420 nulls, and so for each use of the Romano-Wolf test we would expect to falsely reject only 0.05 of a measure. Of course, we run this test on 31 different assets, and so across all assets we expect one or two of the testing procedures to result in false rejections.

Table 5
Number of estimators that are significantly different from RV5min in Romano–Wolf Tests.

QV Proxy:	Worse					Better					Total estimators
	RV daily	*RV* 15 min	*RV* 5 min	*MSRV* 1 min	*RKth2* 1 min	*RV* daily	*RV* 15 min	*RV* 5 min	*MSRV* 1 min	*RKth2* 1 min	
KO	194	243	228	252	249	0	0	0	0	0	418
LSI	183	281	274	288	294	0	0	0	0	0	417
MSFT	284	298	287	302	304	0	0	0	0	0	418
IFF	148	252	268	272	265	0	0	0	0	0	413
SYY	155	225	221	203	203	0	0	0	0	0	414
DGE	184	336	354	244	261	0	0	0	0	0	420
VOD	219	294	371	220	222	0	0	0	0	0	419
SAB	146	338	295	326	329	0	0	0	0	0	420
SDR	142	319	313	277	288	0	0	0	0	0	416
RSA	162	308	381	175	213	0	0	0	0	0	419
TU	246	191	208	179	200	0	0	0	0	0	419
FV	231	254	237	238	253	0	0	0	0	0	420
TY	224	246	230	227	241	0	9	24	28	23	420
US	245	263	257	260	272	0	0	0	0	0	419
FGBL	220	289	286	287	288	0	0	0	0	0	420
FGBS	372	386	143	379	359	0	0	0	0	0	420
CD	141	189	191	190	191	0	0	0	0	0	420
AD	126	183	186	192	193	0	0	0	0	0	420
BP	161	178	182	177	178	0	0	0	0	0	420
URO	177	179	184	184	184	0	0	0	0	0	420
JY	163	185	191	188	185	0	0	0	0	0	420
STXE	211	68	198	299	302	0	0	0	0	0	420
JNI	296	339	348	332	333	0	0	0	0	0	416
FDX	169	157	157	194	193	0	0	0	0	0	420
FFI	175	196	194	196	197	0	0	0	0	0	420
ES	186	216	216	216	218	0	0	0	0	0	420
SPX	182	178	178	161	172	0	0	0	7	1	210
STOXX50E	144	181	179	149	177	0	0	0	0	0	210
DAX	145	157	164	155	161	0	0	0	0	0	210
FTSE	184	186	183	135	180	0	0	0	0	0	210
N225	168	168	170	170	169	0	0	0	0	0	208

Note: Results are displayed shaded gray if the measure used as the QV proxy has a significantly different mean than RVdaily.

RVdaily, an indication of bias, have results shaded gray in Table 5.) Results from the more precise proxies are very similar: we can reject over half of the competing estimators as being significantly worse than RV5min, but we find just one asset out of 31 with *any* measures that significantly outperform RV5min.[26] It is worth noting here that Table 5 reveals that the use of a particular realized measure as a proxy does *not* lead to an apparent improvement in the performance of measures from the same class. Specifically, using a RV as the proxy does not "favor" RV measures, and using RK or TSRV does not favor RK or TSRV measures. The use of a one-day lead of the proxy solves this potential problem.[27]

The asset for which we find that RV5min is significantly beaten, the 10-year US Treasury note futures contract (TY), is among the most frequently traded in our sample. (It is noteworthy, however, that there are five other assets that are even more frequently traded, see Table 1, but for which we find *no* realized measure significantly better than RV5min.) For the 10-year Treasury note, the realized measures that outperform RV5min include MSRV, RK and RRV all estimated using 1-second or 5-second sampling (in calendar-time or business-time, with or without subsampling), and 1-minute RV and 1-minute RVac1; a collection of measures that one might expect to do well for a very liquid asset.

It is also noteworthy, that, combining the set of estimators that are significantly worse than RV5min (around one half of all estimators) with those that are significantly better (approximately zero), leaves, obviously, around one half of the set of 420 estimators that are not significantly different than RV5min in terms of average accuracy.

[26] We also implemented the Romano-Wolf procedure swapping the "reality check" step with a step based on the test of Hansen (2005). This latter test is designed to be less sensitive to poor alternatives with large variances (a potential concern in our application) and so should have better power. We found no change in the number of rejections. In a more forceful attempt to examine the sensitivity to poor alternatives: we identified, ex ante, 72 estimators that the existing literature would suggest are likely to have poor performance (for example, realized kernels on 15-minute returns). We removed this group of estimators from the competing set, and conducted the Romano-Wolf procedure on the remainder of the competing set. We found virtually no change in results of the tests—in fact, counting across the two Romano–Wolf tests for each of 31 assets, there was only one instance where an estimator was found to have different outcome from the original test.

[27] In Table A8 of the web appendix we present some variations of the methods used to obtain the results in Table 5. First, we change the loss function from QLIKE to MSE. Simulation results in Patton and Sheppard (2009a), and empirical results in Hansen and Lunde (2005), Patton and Sheppard (2009b) and Patton (2011a) suggest that tests based on MSE have lower power than under QLIKE, and our results are consistent with this: under MSE we continue to find *no* cases where RV5min is significantly beaten by an alternative, and a reduction in the number of cases where RV5min is identified as significantly better than an alternative. Second we consider this test based on a mean-reverting AR(1) approximation for QV rather than the maintained random walk approximation. Consistent with the simulation results in Patton (2011a) and the results under MSE, we find similar results to our base case but with lower power. Finally we consider the use of a one-day lag, or an average of the one-day lag and one-day lead, as the proxy rather than the one-day lead. The results are mostly unchanged, indicating that either of these instruments for the day t proxy are also suitable.

Table 6
Proportion of realized measures significantly worse than RV5min.

All 31 Assets

	1t	1s	5s	1m	5m	15m
RV	70	57	42	18	16	72
RVss		57	42	18	18	68
RVpa	18	16	26	81	99	99
RVac1	30	45	30	19	49	73
RK	11	15	18	50	87	91
M/TSRV	48	36	44	70	95	91
MLRV	28	43	25	22	84	78
RRV	25	35	27	22	66	94
BR	18					

Interest Rate Futures

	1t	1s	5s	1m	5m	15m
RV	75	58	33	8	50	100
RVss		58	33	8	54	100
RVpa	33	17	92	100	100	100
RVac1	36	44	19	46	81	96
RK	40	18	55	98	99	97
M/TSRV	42	17	33	98	99	92
MLRV	33	38	17	83	100	85
RRV	13	19	13	50	100	98
BR	31					

Index Futures

	1t	1s	5s	1m	5m	15m
RV	60	45	30	0	30	100
RVss		45	30	0	15	90
RVpa	10	5	20	90	100	100
RVac1	10	35	15	10	55	83
RK	10	3	10	64	100	95
M/TSRV	55	30	50	79	99	95
MLRV	20	30	5	15	93	95
RRV	25	28	15	16	79	100
BR	5					

Individual Equities

	1t	1s	5s	1m	5m	15m
RV	65	60	55	25	0	43
RVss		60	55	25	0	35
RVpa	10	13	0	55	98	98
RVac1	40	55	43	14	29	59
RK	0	14	3	28	71	86
M/TSRV	50	49	39	52	90	89
MLRV	30	55	35	0	71	61
RRV	25	50	36	5	39	85
BR	11					

Currency Futures

	1t	1s	5s	1m	5m	15m
RV	70	40	10	0	0	65
RVss		40	10	0	0	65
RVpa	0	0	0	90	100	100
RVac1	0	10	0	0	40	57
RK	0	0	1	41	93	88
M/TSRV	30	10	35	63	98	89
MLRV	0	10	0	0	80	80
RRV	0	0	0	1	57	100
BR	0					

Computed Indices

	1t	1s	5s	1m	5m	15m
RV	100	100	100	80	0	80
RVss		100	100	80	40	90
RVpa	60	80	40	100	100	100
RVac1	80	100	100	30	60	85
RK	5	65	45	18	80	98
M/TSRV	80	100	90	75	98	97
MLRV	80	100	100	20	85	85
RRV	100	100	100	75	80	100
BR	75					

Notes: This table aggregates, for groups of assets (either all 31 assets or assets belonging to one class), the Romano–Wolf test results identifying estimators that are significantly worse than the benchmark 5-minute RV (calendar-time, trades prices) estimator. Each table cell reports the proportion of estimators of a certain estimator class and sampling frequency (across assets, and allowing for different sampling schemes and sampled price series) that are found to be significantly worse than the benchmark estimator in a Romano–Wolf test.

To better understand the results of the Romano-Wolf tests applied to this large collection of assets and realized measures, Table 6 presents the proportion (across assets) of estimators that are significantly worse than RV5min by class of estimator and sampling frequency.[28] Darker shaded regions represent "better" estimators, in the sense that they are rejected less often. Across the five asset classes and the entire set of assets, we observe a darker region running from the top right to the bottom left. This indicates that the simpler estimators in the top rows (RV and variants) do better, on average, when implemented on lower frequency data, such as 1-minute and 5-minute data, while the more sophisticated estimators (RK, MSRV, TSRV and RRV) do relatively better when implemented on higher frequency data, such as 1-second and 5-second data.

5.4. Estimating the set of best realized measures

The tests in the previous section compare a set of competing realized measures with a given benchmark measure. The RV5min measure is a reasonable, widely-used, benchmark estimator, but one might also be interested in determining whether maintaining that estimator as the "null" gives it undue preferential treatment. To address this question, we undertake an analysis based on the "model confidence set" (MCS) of Hansen et al. (2011). Given a set of competing realized measures, this approach identifies a subset that contains the unknown best estimator with some specified level of confidence, with the other measures in the MCS being not significantly different from the true best realized measure. As above, we use the QLIKE distance and a one-day lead of RVdaily as the proxy for QV, and Politis and Romano's (1994) stationary bootstrap with 1000 bootstrap replications and average block-size equal to 10.[29]

The number of realized measures in the model confidence sets varies across individual assets, from 2 to 148 (corresponding to a range of 1% to 36% of all measures), with the average size being 39 estimators, representing just under 10% of our set of 420 realized measures. Index futures and interest rate futures have the smallest model confidence sets, containing around 5% of all realized measures, while individual equities have the largest sets, containing around 17% of all measures. Table A7 in the appendix contains further information on the MCS size for each asset.

In Table 7, we summarize these results by reporting the proportion of estimators from a given class and given frequency that are included in model confidence sets, aggregating results across assets. Darker shaded elements represent the "better" realized measures. Table 7 reveals a number of interesting features. Focusing on the results for all 31 assets, presented in the upper-left panel, we see that the "best" realized measure, in terms of number of appearances in a MCS, is not 5-minute RV but 1-minute subsampled RV. Realized kernels sampled at the one-second frequency also do very well, as do the preaveraged realized variance estimators. The performance of these noise robust measures is particularly strong for individual equities, possibly reflecting this asset class' position as the focus of most existing empirical work.

Looking across asset classes, we see a similar pattern to that in Table 6: a dark region of good estimators includes RV and variants based on lower frequency data (5 s to 5 min) and more sophisticated estimators (RK, MSRV/TSRV, MLRV and RRV) based on higher frequency data (1 s and 5 s). We also observe that for

[28] In this table we aggregate across calendar-time and tick-time, trade prices and quote prices, to focus on the class of realized measure and sampling frequency dimensions.

[29] Similar to above, we also consider 15-minute RV, 5-minute RV, 1-minute MSRV, and 1-minute RKth2 (calendar-time, trades prices) as proxies for QV. Again, we find that using one of these more accurate proxies leads to greater power in the test, i.e. smaller model confidence sets. However, the results show similar patterns to those using RVdaily as the proxy, and importantly, we find that using a proxy of a certain class (RV, TSRV, RK) does not bias the results of the test in favor of estimators of the same class. Detailed results can be found in the online appendix.

Table 7
Proportion of realized measures in 90% model confidence sets.

All 31 Assets

	1t	1s	5s	1m	5m	15m
RV	4	4	14	28	15	0
RVss		4	14	31	19	2
RVpa	7	27	14	0	0	0
RVac1	7	7	18	27	7	1
RK	18	29	26	6	0	0
M/TSRV	4	17	14	1	0	0
MLRV	9	13	23	16	0	0
RRV	14	11	17	19	2	0
BR	10					

Interest Rate Futures

	1t	1s	5s	1m	5m	15m
RV	0	0	0	17	8	0
RVss		0	0	21	8	0
RVpa	0	8	0	0	0	0
RVac1	0	0	8	8	2	0
RK	0	10	2	0	0	0
M/TSRV	0	21	13	0	0	0
MLRV	0	0	23	0	0	0
RRV	4	10	22	6	0	0
BR	0					

Index Futures

	1t	1s	5s	1m	5m	15m
RV	0	0	30	25	0	0
RVss		0	30	25	0	0
RVpa	0	5	0	0	0	0
RVac1	0	3	20	10	0	0
RK	0	10	3	0	0	0
M/TSRV	0	13	4	0	0	0
MLRV	0	20	18	0	0	0
RRV	15	8	18	0	0	0
BR	5					

Individual Equities

	1t	1s	5s	1m	5m	15m
RV	10	10	13	45	23	0
RVss		10	13	53	38	5
RVpa	15	60	33	0	0	0
RVac1	20	13	21	46	15	3
RK	33	58	55	4	0	0
M/TSRV	10	15	28	0	0	0
MLRV	25	13	25	30	0	0
RRV	28	10	18	44	5	0
BR	19					

Currency Futures

	1t	1s	5s	1m	5m	15m
RV	0	0	25	20	10	0
RVss		5	25	20	20	0
RVpa	10	20	15	0	0	0
RVac1	0	13	33	15	8	0
RK	10	20	18	1	0	0
M/TSRV	0	30	8	3	0	0
MLRV	0	30	35	10	0	0
RRV	5	20	18	13	0	0
BR	13					

Computed Indices

	1t	1s	5s	1m	5m	15m
RV	0	0	0	10	40	0
RVss		0	0	0	10	0
RVpa	0	0	0	0	0	0
RVac1	0	0	0	50	5	0
RK	50	10	28	48	0	0
M/TSRV	10	0	0	10	0	0
MLRV	0	0	0	40	0	0
RRV	0	0	0	0	5	0
BR	0					

Notes: This table aggregates, for groups of assets (either all 31 assets or assets belonging to one class), the 90% Model Confidence Sets identifying the subset containing "best" estimators. Each table cell reports the percentage of all estimators of a certain estimator class and sampling frequency (across assets, and aggregating estimators using different sampling schemes and sampled price series) that are found to be in a Model Confidence Set.

more liquid asset classes, such as currency futures, interest rate futures, and index futures, realized measures appear in a MCS more often if based on higher frequency data. In contrast, for individual equities and for computed equity indices, the preferred sampling frequencies are generally lower.

We can also use the estimated model confidence sets to shed light on the particularly poorly performing realized measures. Across all 31 assets, we see that realized measures based on 15-minute data almost never appear in a MCS. Similarly, we observe that the more sophisticated realized measures, TSRV/MSRV, MLRV, RK and RRV are almost never in a MCS when sampling every 5- or 15-minutes, which appears to be too low for these estimators. In addition, for the highly liquid futures contracts, even a sampling frequency of 1-minute is not high enough to use with these sophisticated realized measures. This is consistent with the implementations of these estimators in the papers that introduced them to the literature, and so is not surprising.

Overall, the results from the previous section revealed that it was very rare to find a realized measure that significantly outperformed 5-minute RV. The analysis in this section, which avoids the need to specify a "benchmark" realized measure, reveals evidence that some measures are indeed more accurate than 5-minute RV. We find that 1-minute RV and RVac1, 1-second and 5-second Realized Kernels and Multi-scale RV, and 5-second and 1-minute Realized Range estimators appear more often in the MCS than 5-minute RV. Subsampled RV at moderate frequencies (1-minute or 5-minute) also outperforms regular 5-minute RV.

5.5. Explaining performance differences

We now seek to shed light on the factors that explain the differences in the accuracy of the realized measures considered in this paper. The results in the previous sections are related to average accuracy over the entire sample period, and in this section we study relative conditional accuracy using a panel version of the approach of Giacomini and White (2006). This approach enables us to examine whether the relative performance of two realized measures varies with some set of conditioning variables. We use a panel specification to exploit both the time series and the cross-sectional information in our data, and we consider a variety of conditioning variables to try to explain when one measure outperforms another. All of the specifications that we consider are of the form:

$$L(\tilde{\theta}_t^i, M_{0,t}^i) - L(\tilde{\theta}_t^i, M_{j,t}^i) = \boldsymbol{\beta}_j' \mathbf{X}_{t-1}^i + \gamma_j' \mathbf{Z}^i + \varepsilon_{j,t}^i,$$
$$\text{for } t = 1, 2, \ldots, T;\ i = 1, 2, \ldots, N \tag{8}$$

where the first realized measure, $M_{0,t}^i$, is taken to be RV5min, and the competing measure, $M_{j,t}^i$, is one of the better-performing realized measures identified in the previous section, namely, 5-second MSRV, 1-minute RVac1, 5-second RKth2 and 1-second MLRV.[30] We also include 1-minute RV and RVdaily to study the accuracy gains from using higher-frequency price data. All of these estimators are computed on transaction prices with calendar-time sampling. The panel is unbalanced as the assets do not all trade on the same days, and the maximum dimensions of the panel are $T = 2860$ and $N = 31$. We estimate this model using an unbalanced panel framework and Driscoll and Kraay (1998) standard errors, which are robust to heteroskedasticity, serial correlation, and cross-sectional dependence.[31]

Our conditioning variables include a variety of variables $\left(\mathbf{X}_{t-1}^i\right)$ that might be thought to influence the accuracy of realized measures. Numerous measures of market liquidity exist in the literature (see Hautsch and Podolskij (2013) and Aït-Sahalia and

[30] The fact that we examine realized measures identified as "good" in previous analysis of course biases the interpretation of any subsequent tests of *unconditional* accuracy. In this section we focus on whether the relative performance of these measures varies significantly with some conditioning variables (X, Z), and the problem of pre-test bias does not arise here.

[31] Panel regressions were estimated using the Stata program "xtscc" (Hoechle, 2007) downloaded from http://ideas.repec.org/c/boc/bocode/s456787.html. These standard errors are essentially a "HAC of cross-sectional averages", and based on the length of the data, the program selects 8 lags to use for the Newey–West kernel. We also used an alternate version of Driscoll–Kraay standard errors developed in Vogelsang (2012), which uses fixed-b asymptotic theory, but we found that the differences in the standard errors were extremely small in this application.

Yu (2009), for examples). We use three measures (see Bandi and Russell (2006) and Diebold and Strasser (2013), for example) that can be computed each day using our intraday price and volume data: average time between trades (avgdur); total trade volume in units of local currency (volm); and average bid–ask spreads (BAsprd). All three of these variables exhibit a strong trend over our sample period, and so we de-trend each series using a 60-day moving average.[32] Next we look at measures of noise and jumps in the asset price series. For each day we measure the autocorrelation in 5-second returns (ac1_5s). The autocorrelations of sampled intraday returns have been studied in the context of measuring QV (Hansen and Lunde, 2006b; Hautsch and Podolskij, 2013) because they embed information about the properties of microstructure noise, which may influence the performance of realized measures. We also include an estimate of relative size of the noise: the per-trade ratio of the noise variance to the total variance (noiseratio), introduced in Oomen (2006a) and used by Hautsch and Podolskij (2013),[33] and the proportion of QV attributable to jumps (jumpprop), measured as $\max\left(RV_t - BPV_t, 0\right)/RV_t$ (both measures use 5-minute tick-time sampled trades prices), to see if the magnitude of noise or jump activity has an effect on realized measure performance. Finally, we include a measure of volatility (logQV), and for this we use subsampled RV5min. Tables A9 and A10 in the web appendix present some summary statistics on these conditioning variables, including the full-sample averages by which the variables are de-meaned, and information on their cross-correlations.[34,35]

Given the panel nature of this analysis, we are also able to include variables $\left(\mathbf{Z}^i\right)$ to capture some of the cross-sectional variation in our data. We first include dummy variables for each asset class (equities, bond futures, FX futures, index futures, and computed indices), which capture some of the time-invariant features of the markets on which these assets trade.[36] We also include geographic dummy variables (US, UK, Europe and Asia, with the US dummy dropped to avoid perfect multicollinearity) to see whether there are differences across countries in the relative performance of these realized measures.[37]

Table 8 presents the results of these panel regressions, with each column of this table representing a separate estimation to compare RV5min with the competing measure listed in the top row. We present the t-statistics above and the parameter estimates in parentheses below. We focus on the t-statistics since, like Diebold–Mariano-type tests, the average loss differential, or in this case, the conditional loss differential, is difficult to directly interpret. The t-statistics in the middle panel correspond to the coefficients on the asset class dummies, and can be interpreted as those on the average difference in performance holding the conditioning variables at their average levels. The first column of Table 8 confirms our earlier results: RVdaily is significantly worse that RV5min across all asset classes, evidenced by the large and negative t-statistics for all four asset class dummy variables. For most of the other, more sophisticated estimators, we find that RV5min is significantly beaten, on average, particularly for the very liquid FX futures and index futures asset classes.

Looking at the conditioning variables across the columns we see that two variables in particular exhibit power in explaining differences in the performance of realized measures. The first of these (noiseratio) measures the variance of the microstructure noise relative to the variance of the return. We see a consistent implication from the coefficients on this variable that as microstructure noise increases, the performance of the more sophisticated realized measure deteriorates. In the first column the coefficient is positive, indicating that the performance of RVdaily improves relative to that of RV5min as the noise increases, while in all of the remaining columns, which consider estimators that are in some way more sophisticated than RV5min, the coefficient is negative. This result is in line with intuition: more sophisticated estimators, sampled at higher frequencies, are more exposed to microstructure noise than less sophisticated alternatives, and so as the level of noise increases we would expect the former to perform *relatively* worse than the latter. Given the inclusion of an intercept in this specification (the asset class dummy variables) this does not necessarily imply as the level of noise increases the researcher should switch from a more sophisticated estimator to a less sophisticated, lower frequency, estimator; rather this just suggests that the gains from using a more sophisticated estimator fall in those circumstances.[38]

The second conditioning variable with substantial predictive power for relative performance is the level of volatility (logQV). The estimation error of volatility measures is generally increasing with the level of volatility, which may explain the usefulness of this variable. The coefficient on this variable is generally negative and significant, which has a different interpretation for the first column than the remaining. In the first column, the negative coefficient suggests that as volatility rises, the gains to using RV5min increase. That is, in high volatility periods the gains to using a moderately high sampling frequency rather than daily sampling are particularly large. In the remaining columns, a negative coefficient indicates that the gains from using a more sophisticated realized measure rather than RV5min *decrease* in periods of high volatility. This may be because the performance of the more sophisticated estimators deteriorates faster than that of RV5min as volatility increases, or it may be related to the fact that the level of volatility is positively correlated with measures of market Illiquidity such as the bid–ask spread and 5-second return correlations, both of which are *ex ante* expected to indicate worse conditions for more sophisticated estimators.

[32] Specifically, we set $\tilde{X}_t = X_t/\bar{X}_{t-60,t-1}$ where $\bar{X}_{t-60,t-1}$ is the average value of the variable over the past 60 days. We also examined using 120- and 250-day averages and the results were qualitatively similar.

[33] We compute noiseratio as in Hautsch and Podolskij (2013): $\text{noiseratio}_t = \hat{\alpha}_t/(\hat{QV}_t/n_t)$, where the numerator is the first-order autocorrelation of the tick-by-tick returns, and the denominator is 1-tick MLRV scaled by the number of observations on that day.

[34] We winsorize all conditioning variables at the 0.5% and 99.5% levels to reduce the impact of extreme observations, and we de-mean all conditioning variables so that the coefficients on the dummy variables can be interpreted as the average loss difference when all conditioning variables are at their full-sample average values. Finally, we normalize the loss differences by their full-sample standard deviations so that the parameter estimates in Table 8 are comparable across columns.

[35] The two strongest correlations are between avgdur and volm at −0.66, and ac1_5s and noiseratio at −0.33. logQV is also mildly correlated to all of the conditioning variables except jumpprop, with correlation magnitudes ranging from 0.22 to 0.32. The remaining cross-correlations are fairly small.

[36] We do not have measures of volume and bid–ask spreads for our computed indices due to the fact that they are constructed series rather than traded assets, and so these two conditioning variables are missing. In the interests of retaining these two interesting variables, we drop the five computed indices from the panel specification reported in Table 8, thereby reducing the cross-sectional dimension to 26. In the web appendix (Table A11), we report a corresponding table with the computed indices included in the panel and with average spread and volume variables dropped. The main conclusions from this sub-section are unaffected.

[37] Note that the "Asia" dummy variable is dominated by Japan: this group only includes the Nikkei 225 index future, the Nikkei 225 computed index, the yen/USD exchange rate and the Australian dollar/USD exchange rate. Further, since all currency pairs are against the USD, we assign the currency futures to the geographic region of its pairing, with the exception of the CAD/USD currency futures, which are assigned to the "US" (in effect, "North America") region.

[38] To determine whether the researcher should switch one would need to zoom in on the loss difference for a given pair of estimators, for a given asset, on a given day. Our panel specification allows for such analyses, but in the interests of space and generality we do not attempt this here.

Table 8
Conditional Relative Performance of Realized Measures and RV5min.

RV5m vs.	**RVdaily**	**RV 1m**	**MSRV 5s**	**RVac1 1m**	**RKth2 5s**	**MLRV 1s**
avgdur	0.74	0.10	-0.44	-1.72	-1.57	1.28
	(0.01)	(0.00)	(-0.01)	(-0.02)	(-0.05)	(0.02)
volm	1.55	-1.74	-1.38	-0.36	-1.24	-1.97
	(0.01)	(-0.01)	(-0.01)	(0.00)	(-0.01)	(-0.01)
BAsprd	1.54	-1.57	-1.23	-1.78	-1.87	-2.57
	(0.06)	(-0.04)	(-0.03)	(-0.03)	(-0.06)	(-0.06)
ac1_5s	-0.70	-4.21	-1.26	-1.63	-0.56	1.65
	(-0.05)	(-0.41)	(-0.10)	(-0.21)	(-0.04)	(0.15)
jumpprop	0.97	0.04	0.28	-0.40	1.33	-2.17
	(0.03)	(0.00)	(0.01)	(-0.02)	(0.05)	(-0.08)
noiseratio	3.56	-8.84	-4.98	-2.71	-2.14	-13.11
	(0.01)	(-0.05)	(-0.02)	(-0.03)	(-0.01)	(-0.06)
logQV	-4.94	-2.02	-2.62	1.85	-2.11	0.91
	(-0.03)	(-0.01)	(-0.02)	(0.01)	(-0.01)	(0.01)
equities	-12.38	-0.63	0.60	0.38	4.23	-12.68
	(-0.12)	(0.00)	(0.00)	(0.00)	(0.03)	(-0.08)
bond fut	-10.26	6.34	5.90	-4.27	4.52	-0.80
	(-0.09)	(0.17)	(0.14)	(-0.06)	(0.08)	(-0.02)
FX fut	-8.59	7.39	5.74	1.20	5.94	6.71
	(-0.16)	(0.10)	(0.07)	(0.02)	(0.07)	(0.09)
index fut	-8.79	10.45	9.00	-0.47	6.79	4.49
	(-0.09)	(0.13)	(0.09)	(0.00)	(0.08)	(0.06)
UK	-0.57	-18.11	-12.96	-3.08	-7.74	-21.47
	(-0.01)	(-0.19)	(-0.11)	(-0.03)	(-0.07)	(-0.17)
Europe	2.28	-8.31	-6.47	0.09	-4.08	-11.90
	(0.03)	(-0.12)	(-0.08)	(0.00)	(-0.05)	(-0.17)
Asia	-0.96	-7.19	-3.54	-1.52	-1.76	-16.70
	(-0.02)	(-0.11)	(-0.04)	(-0.05)	(-0.02)	(-0.24)

Notes: Each column of this table presents the t-statistics (top) and coefficient estimates (bottom, in parentheses) for a pooled regression of the form $L(\tilde{\theta}_t^i, M_{0,t}^i) - L(\tilde{\theta}_t^i, M_{j,t}^i) = \beta_j' \mathbf{X}_{t-1}^i + \gamma_j' \mathbf{Z}^i + \varepsilon_{j,t}^i$, for $t = 1, 2, \ldots, T$; $i = 1, 2, \ldots, 26$, where $M_{0,t}^i$ is RV5min, $M_{j,t}^i$ is a competing realized measure listed in the table header, $\mathbf{X}_{t-1}^i$ are the set of 7 explanatory variables listed in the first 7 rows of first column, and $\mathbf{Z}^i$ are the set of 7 categorical variables listed in the last 7 rows of the first column. 26 assets (all assets other than the computed indices) are included in each panel regression, and T=2860 (though panel is unbalanced). All estimators are calendar-time sampled, transaction price estimators. Statistically significant results (at 5% level) are shaded.

Finally we turn to the bottom panel of Table 8, which presents the coefficients and t-statistics on geographic dummy variables. These coefficients represent the average loss difference between two estimators across all assets in a given geographic region, compared with those for the assets in the US. The main result that emerges from this panel, based on columns 2 to 6, is that the more sophisticated estimators tend to perform better (relative to RV5min) for US assets than for European or Asian assets; the parameter estimates are almost all negative and significant. This is perhaps related to the fact that much of the work underlying these estimators has been undertaken with US assets in mind.

This analysis sheds light on the market conditions and market features that lead to variations in the relative performance of measures of volatility. Sophisticated estimators, such as MSRV, RK, and RVpa, which perform well on average, see their performance deteriorate relative to the simple 5-minute RV estimator in periods of high noise and high volatility, periods that may be particularly important to investors since they can represent conditions of crisis (low liquidity and high volatility, as in the recent financial crisis). In such times accurate measurement of volatility is important for both pricing and risk management. The results from this section also reveal the importance of US equity markets in the development of sophisticated estimators. These estimators generally perform significantly worse in non-US markets than in US markets. This is perhaps indicative of microstructure effects of a different form to those in US markets, providing motivation for more detailed analyses of non-US market microstructures.

5.6. Out-of-sample forecasting with realized measures

The results above have all focused on the relative accuracy of realized measures for *estimating* quadratic variation. One of the main uses of estimators of volatility is in the production of volatility forecasts, and in this section we compare the relative accuracy of forecasts based on our set of competing realized measures. We do so based on the simple heterogeneous autoregressive (HAR) forecasting model of Corsi (2009). This model is popular in practice as it captures long memory-type properties of quadratic variation, while being simpler to estimate than fractionally integrated processes, and performs well in volatility forecasting, see Andersen et al. (2007) for example. For each realized measure, we estimate the HAR model using the most recent 500 days of data:

$$\tilde{\theta}_{t+h} = \beta_{0,j,h} + \beta_{1,j,h} M_{jt} + \beta_{2,j,h} \frac{1}{5} \sum_{k=0}^{4} M_{j,t-k} + \beta_{3,j,h} \frac{1}{22} \sum_{k=0}^{21} M_{j,t-k} + \varepsilon_{jt}, \tag{9}$$

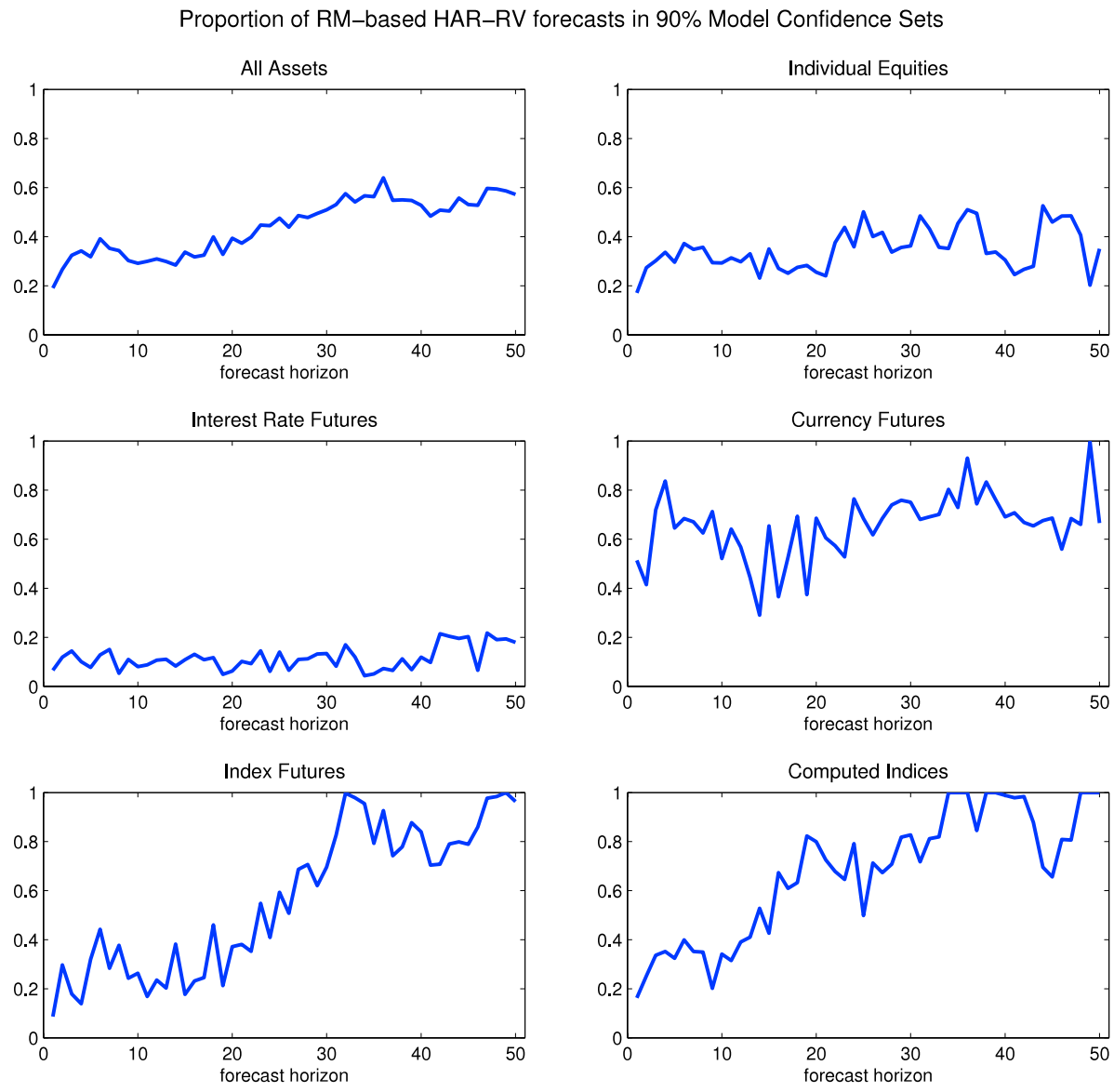

Fig. 1. This figure presents the proportion of all realized measure based HAR-RV forecasts of QV that are included in the 90% model confidence set at each forecast horizon, ranging from 1 to 50 days. The upper left panel presents the results over all 31 assets, and the remaining panels present results for each of the five asset classes separately.

where M_{jt} is a realized measure from the competing set, and $\tilde{\theta}_{t+h}$ is the volatility proxy (the squared open-to-close return for day $t+h$, which is a proxy for QV). We estimate this regression separately for each forecast horizon, h, ranging from 1 to 50 trading days, and from those estimates we obtain a h-day ahead volatility forecast, which we then compare with our volatility proxy. We re-estimate the model each day using a rolling window of 500 days.

In addition to the 420 realized measures we have analyzed so far, for forecasting analysis we also consider some "jump-robust" estimators of volatility. These measures, described in Section 2.3, are designed to estimate only the integrated variance component of quadratic variation, see Eq. (2). The inclusion of these estimators is motivated by studies such as Andersen et al. (2007) and Patton and Sheppard (forthcoming), which report that the predictability of the integrated variance component of quadratic variation is stronger than the jump component, and thus there may be gains to separately forecasting the two components. Using a HAR model on these jump-robust realized measures effectively treats the jump component as unpredictable, while using a HAR model on estimators of QV (our original set of 420 measures) treats the two components as having equal predictability.[39] (These are of course extreme viewpoints; a more nuanced approach would allow both components to have non-zero and possibly different levels of predictability, as in Andersen et al. (2007), but in the interests of space we do not consider that here.) Extending our set to include 228 jump-robust measures increases its total number to 648 realized measures.

For each forecast horizon between one day and 50 days we estimate the model confidence set of Hansen et al. (2011). It is not feasible to report the results of each of these estimates for each horizon, and so we summarize them in two ways. Firstly, in Fig. 1 we present the size of the MCS, measured as the proportion

[39] If QV is comprised of two AR(1) components (namely integrated variation and jump variation) with differing degrees of persistence, then it will follow an ARMA(1,1) process. This is clearly not consistent with our maintained random walk approximation for QV. In Table A8 of the web appendix we consider two alternative approximations for QV, an AR(1) and an AR(5), the latter motivated as an alternative to an ARMA approximation. We find reduced power from these tests, but the rejections we do obtain are consistent with those found under the random walk approximation.

Table 9
Proportion of RM-based HAR-RV models in 90% Model Confidence Sets, for forecast horizons 1 through 5.

All 31 Assets

	1t	1s	5s	1m	5m	15m
RV	12	14	17	25	35	30
RVss		16	16	25	35	44
RVpa	40	34	36	41	23	15
RVac1	14	18	17	29	33	27
RK	33	30	34	35	30	13
M/TSRV	19	17	19	31	28	11
MLRV	19	18	20	33	34	11
RRV	18	15	17	37	42	27
BR	36					
BPV	10	13	16	25	42	49
BPVpa	42	34	47	60	41	23
min/medRV	11	13	15	23	38	44
QRV	11	12	16	36	63	55
TrunRV	11	11	22	42	61	63

Interest Rate Futures

	1t	1s	5s	1m	5m	15m
RV	2	3	8	0	1	3
RVss		5	7	1	0	0
RVpa	0	0	0	4	3	8
RVac1	2	5	6	0	3	1
RK	0	0	1	5	6	2
M/TSRV	3	7	5	0	5	4
MLRV	2	8	5	0	6	0
RRV	2	5	2	0	9	13
BR	2					
BPV	0	0	1	2	20	34
BPVpa	10	11	37	53	49	30
min/medRV	0	0	1	3	22	29
QRV	0	0	6	39	74	78
TrunRV	0	0	4	57	72	88

Index Futures

	1t	1s	5s	1m	5m	15m
RV	18	16	16	17	25	16
RVss		17	17	17	25	33
RVpa	36	30	24	32	12	5
RVac1	10	18	13	19	24	17
RK	28	22	29	26	21	3
M/TSRV	19	14	15	20	20	3
MLRV	26	18	20	24	19	6
RRV	22	15	13	27	21	10
BR	47					
BPV	4	2	22	18	32	32
BPVpa	30	23	35	39	13	5
min/medRV	5	6	17	16	26	31
QRV	4	13	13	29	44	22
TrunRV	3	1	17	30	54	39

Individual Equities

	1t	1s	5s	1m	5m	15m
RV	5	6	6	23	42	35
RVss		0	5	23	43	58
RVpa	46	35	41	49	32	17
RVac1	12	5	9	31	38	35
RK	40	32	37	39	35	14
M/TSRV	15	11	16	32	31	10
MLRV	15	5	10	35	42	11
RRV	10	4	5	43	58	33
BR	37					
BPV	6	2	3	16	38	49
BPVpa	46	29	38	58	39	18
min/medRV	7	3	3	16	32	43
QRV	6	0	4	13	46	47
TrunRV	7	-	0	16	39	46

Currency Futures

	1t	1s	5s	1m	5m	15m
RV	28	42	47	64	78	71
RVss		40	46	63	79	73
RVpa	72	70	79	69	49	39
RVac1	30	52	49	71	75	47
RK	68	68	74	80	62	32
M/TSRV	38	46	46	77	57	30
MLRV	38	52	49	80	68	33
RRV	44	46	57	81	76	54
BR	64					
BPV	24	43	44	72	90	86
BPVpa	80	85	93	93	68	47
min/medRV	26	38	44	68	84	79
QRV	28	34	52	84	96	86
TrunRV	30	40	53	85	91	96

Computed Indices

	1t	1s	5s	1m	5m	15m
RV	16	16	18	30	20	20
RVss		0	0	34	22	54
RVpa	52	40	38	58	4	0
RVac1	20	13	14	32	20	34
RK	26	29	29	20	24	14
M/TSRV	30	13	14	24	35	4
MLRV	28	13	34	32	34	0
RRV	26	20	35	35	34	16
BR	36					
BPV	32	-	-	34	33	45
BPVpa	56	32	40	68	12	0
min/medRV	32	-	0	19	37	42
QRV	32	0	12	42	78	34
TrunRV	24	-	13	24	74	53

Notes: This table aggregates, for groups of assets (either all 31 assets, or assets belonging to one class), the 90% Model Confidence Sets identifying the subset containing "best" estimators. Each table cell reports the percentage of all estimators of a certain estimator class and sampling frequency (across assets, and aggregating estimators using different sampling schemes and sampled price series) that are found to be in a Model Confidence Set. '-' indicates that for the assets under consideration, all estimators of that class and sampling frequency yield values that are unrealistically small and thus dropped from the competing set (see section 7.2 in the Appendix).

of realized measures that are included in the MCS, across forecast horizons. From this figure we observe that the MCSs are relatively small for short horizons, consistent with our results in Section 5.4 and with the well-known strong persistence in volatility. As the forecast horizon grows, the size of the MCSs increase, reflecting the fact that for longer horizons more precise measurement of current volatility provides less of a gain than for short horizons. It is noteworthy that even at horizons of 50 days, we are able to exclude 43% of realized measures from the MCS, averaging across all 31 assets. This proportion varies across asset classes, with the proportion of estimators included at $h = 50$ equal to 18% for the class of interest rate futures, 35% for individual equities, and near 100% (i.e., no realized measures are excluded) for computed equity indices, index futures and currency futures.

In Table 9 we study these results in greater detail. This table has the same format as Table 7, and reports the proportion of realized measures from a given class and given frequency that belong to a model confidence set, aggregating results across assets and forecast horizons between 1 and 5 days. As in Table 7, darker shaded elements represent the better forecasts. What is most striking about this table is the relative success of the jump-robust realized measures for volatility forecasting. For four of the five asset classes, the best measure is one of truncated RV (TRV) at the 5-minute or 15-minute frequency, or quantile-RV (QRV) at the 5-minute frequency. Individual equities is the only asset class where nonjump-robust estimators (RVss on 15-minute sampling and 5-minute RRV) tie with a noise-robust estimator (1-minute BPVpa) for highest proportion of forecasts belonging in MCSs. This broad pattern that the best realized measures for volatility forecasting appears to be jump-robust measures, estimated using relatively low (5- or 15-minute) frequency data is consistent with the existing results in the literature, see Andersen et al. (2007) and Corsi et al. (2010) for example, who find that separating QV into continuous and jump components leads to better out-of-sample forecast performance.

In Fig. 2 we present the proportion (across assets) of model confidence sets that contain RV5min and TRV5min (both computed on transaction prices with calendar-time sampling), for each forecast horizon. We see that, across all assets, RV5min appears in around 30% of MCSs for shorter horizons, rising to around 60% for longer horizons.[40] RV5min does best for currency futures and individual equities, and relatively poorly for interest rate futures. Fig. 2 also presents the corresponding proportion for TRV5min, and we see that this measure does almost uniformly better than RV5min, with the exceptions being for the individual equities, where it is dominated by RV5min, and index futures, where TRV5min and RV5min forecasting models show similar performance. TRV5min does particularly well for currency futures and interest rate futures.

Our study of a broad collection of assets and a large set of realized measures necessitates simplifying the analysis in several ways, and a few caveats to the above conclusions apply. Firstly, these results are based on each realized measure being used in conjunction with the HAR model of Corsi (2009). This model has proven successful in a variety of volatility applications, but it is by no means the only relevant volatility forecasting model in

[40] Note that this analysis only counts RV5min computed in calendar time, using transaction prices, and not subsampled. Thus this represents a lower bound on the proportion of MCSs that include *any* RV5min.

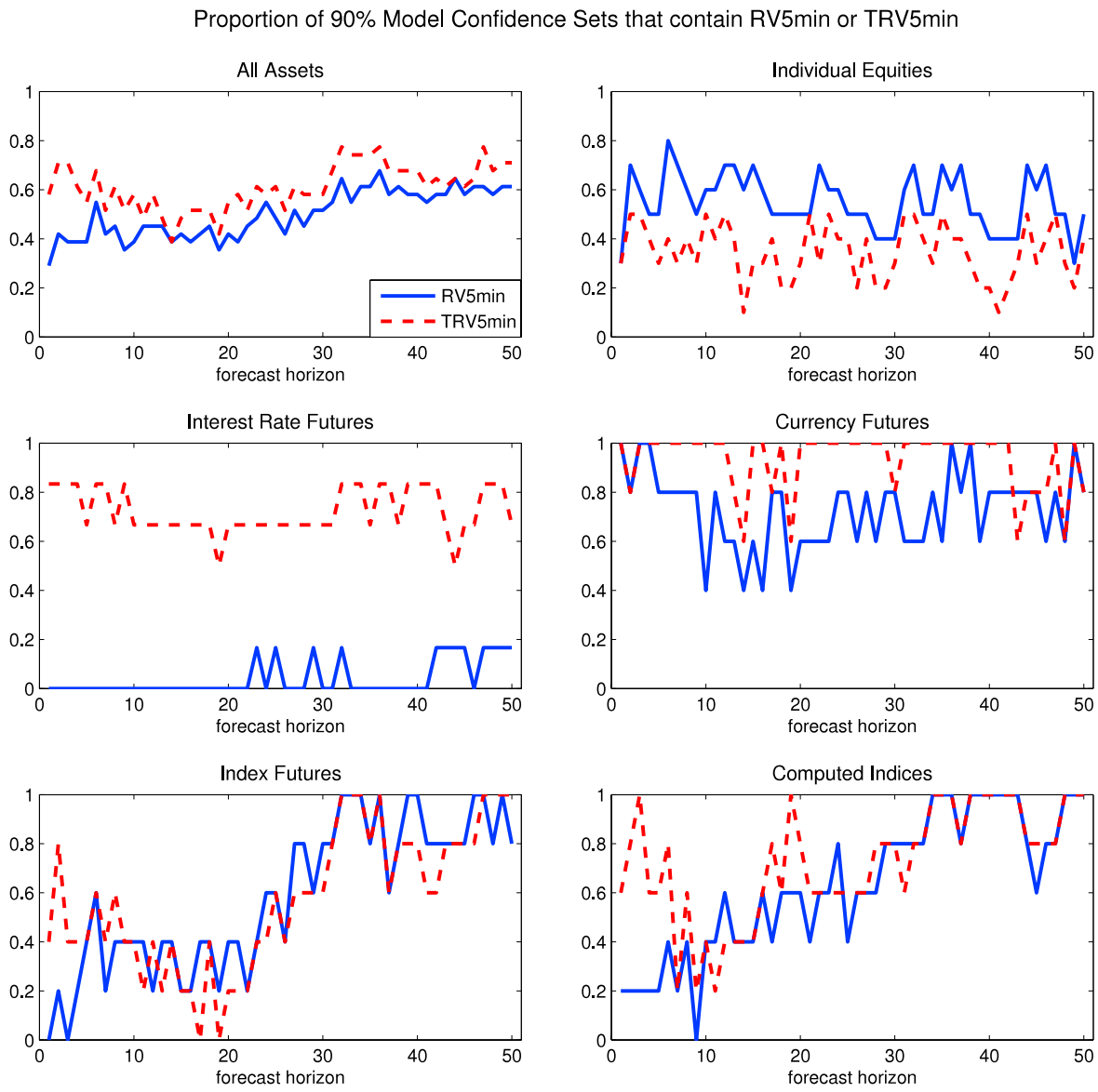

Fig. 2. This figure presents the proportion of 90% model confidence sets (across assets) that contain 5-minute RV and 5-minute truncated RV (under calendar-time sampling, and using transactions prices if available) at each forecast horizon ranging from 1 to 50 days. The upper left panel presents the results across all 31 assets, and the remaining panels present results from each of the 5 asset classes separately.

the literature, and it is possible that the results and rankings change with the use of a different model. Secondly, by treating the prediction of future QV as a univariate problem, we have implicitly made a strong assumption about the predictability of volatility attributable to jumps, either that it is identical to that of integrated variance, or that it is not predictable at all. A more sophisticated approach might treat these two components separately. Thirdly, we have only considered forecasting models based on a single realized measure, and it may be possible that a given realized measure is not very useful on its own, but informative when combined with another realized measure.

6. Summary and conclusion

Motivated by the large body of research on estimators of asset price volatility using high frequency data (so-called "realized measures"), this paper considers the problem of comparing the empirical accuracy of a large collection these measures across a range of assets. In total, we consider over 400 different estimators, applied to 11 years of data on 31 different financial assets across five asset classes, including equities, indices, exchange rates and interest rates. We apply data-based ranking methods to the realized measures and to forecasts based on these measures, for forecast horizons ranging from 1 to 50 trading days.

Our main findings for these 31 assets can be summarized as follows. Firstly, if 5-minute RV is taken as the benchmark realized measure, then using the testing approach of Romano and Wolf (2005), we find very little evidence that it is significantly outperformed by *any* of the competing measures in terms of estimation accuracy, across any of the 31 assets under analysis. If, on the other hand, the researcher wishes to remain agnostic about the "benchmark" realized measure, then using the model confidence set of Hansen et al. (2011), we find that 5-minute RV is indeed outperformed by a small number of estimators, most notably 1-minute sub-sampled RV, and 1- and 5-second realized kernels and MSRV. Finally, when using forecast performance as the method of ranking realized measures, we find that 5-minute or 15-minute truncated RV provides the best performance on average, which is consistent with the work of Andersen et al. (2007), who find that jumps are not very persistent. The rankings of realized

measures vary across asset classes, with 5-minute RV performing better on the relatively less liquid classes (individual equities and computed equity indices), and the gains from more sophisticated estimators like MSRV and realized kernels being more apparent for more liquid asset classes (such as currency futures and equity index futures). We also find that for realized measures based on frequencies of around five minutes, sampling in tick time and subsampling the realized measure both generally lead to increased accuracy.

It is important to acknowledge here that while we consider a relatively large collection of assets, they all share the characteristic of being relatively liquid assets on well-developed markets, and our conclusions require some adjustment before considering them for other assets. We suggest the following three general conclusions. First, sampling relatively sparsely appears to accrue much of the benefits of "high frequency" data (whatever that means for a given asset) without exposing the measure to problems from microstructure noise. Five-minute sampling is an example of sparse sampling for moderately liquid assets; for less liquid assets 15 min to one hour might be more appropriate, and as the assets we study get more liquid, one-minute sampling may be interpretable as "sparse". Second, "subsampling" (Zhang et al., 2005) is an easy, and robust, way to improve the accuracy of sparsely-sampled realized measures. Finally, the gains from high frequency data are greatest when microstructure noise, somehow measured, is relatively low, and when volatility is high. These two quantities can vary substantially through time, as well as across assets. Investigating the performance of these, and newly-developed, realized measures on an even broader set of assets (less liquid, perhaps on developing markets) is an interesting avenue for future research.

Acknowledgments

We thank two referees, an associate editor, the editor (Yacine Aït-Sahalia), and Tim Bollerslev, Jia Li, Asger Lunde, George Tauchen, Aurelio Vasquez, Julian Williams, and seminar participants at Cass Business School, Duke University, ITAM, the conference in honor of Timo Teräsvirta at Aarhus University, the Ultra High Frequency Econometrics workshop on the Isle of Skye, the 2012 Society for Financial Econometrics summer school, and the 2013 Econometric Society summer meetings for helpful comments.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2015.02.008.

References

Aït-Sahalia, Y., Jacod, J., 2014. High-Frequency Financial Econometrics. Princeton University Press, New Jersey.
Aït-Sahalia, Y., Mykland, P.A., Zhang, L., 2005. How often to sample a continuous-time process in the presence of market microstructure noise. Rev. Financial Studies 18, 351–416.
Aït-Sahalia, Y., Yu, J., 2009. High frequency market microstructure noise estimates and liquidity measures. Ann. Appl. Stat. 3, 422–457.
Alizadeh, S., Brandt, M.W., Diebold, F.X., 2002. Range-based estimation of stochastic volatility models. J. Finance 57, 1047–1092.
Andersen, T.G., Bollerslev, T., 1998. Answering the skeptics: yes, standard volatility models do provide accurate forecasts. Internat. Econom. Rev. 39, 885–905.
Andersen, T.G., Bollerslev, T., Christoffersen, P., Diebold, F.X., 2006. Volatility and correlation forecasting. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 1. Elsevier, Oxford.
Andersen, T.G., Bollerslev, T., Diebold, F.X., 2007. Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility. Rev. Econom. Statist. 89, 701–720.
Andersen, T., Bollerslev, T., Diebold, F., Ebens, H., 2001a. The distribution of realized stock return volatility. J. Financial Economics 61 (1), 43–76.
Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001b. The distribution of realized exchange rate volatility. J. Amer. Statist. Assoc. 96 (453), 42–55.
Andersen, T.G., Bollerslev, T., Meddahi, N., 2005. Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. Econometrica 73 (1), 279–296.
Andersen, T., Dobrev, D., Schaumburg, E., 2012. Jump-robust volatility estimation using nearest neighbor truncation. J. Econometrics 169 (1), 75–93.
Bandi, F., Russell, J., 2006. Separating microstructure noise from volatility. J. Financial Economics 79, 655–692.
Bandi, F.M., Russell, J.R., 2008. Microstructure noise, realized variance, and optimal sampling. Rev. Econom. Stud. 75 (2), 339–369.
Barndorff-Nielsen, O.E., 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (2), 253–280.
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 (6), 1481–1536.
Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2011. Subsampling realised kernels. J. Econometrics 160 (1), 204–219.
Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2009. Realized kernels in practice: trades and quotes. Econometrics Journal 12 (3), C1–C32.
Barndorff-Nielsen, O.E., Shephard, N., 2007. Variation, jumps, market frictions and high frequency data in financial econometrics. In: Blundell, R., Torsten, P., Newey, W.K. (Eds.), Advances in Economics and Econometrics. Theory and Applications. In: Econometric Society Monographs, Cambridge University Press, Cambridge, pp. 328–372.
Barndorff-Nielsen, O., Shephard, N., 2006. Econometrics of testing for jumps in financial economics using bipower variation. J. Financial Econometrics 4 (1), 1–30.
Bibinger, M., Hautsch, N., Malec, P., Reiss, M., 2014. Estimating the quadratic covariation matrix from noisy observations: local method of moments and efficiency. Ann. Statist. 42 (4), 80–114.
Bollerslev, T., Engle, R.F., Nelson, D.B., 1994. ARCH models. Handbook of Econometrics 4, 2959–3038.
Christensen, K., Oomen, R.C., Podolskij, M., 2014. Fact or friction: Jumps at ultra high frequency. J. Financial Economics 114 (3), 576–599.
Christensen, K., Oomen, R., Podolskij, M., 2010. Realised quantile-based estimation of the integrated variance. J. Econometrics 159 (1), 74–98.
Christensen, K., Podolskij, M., 2007. Realized range-based estimation of integrated variance. Journal of Econometrics 141 (2), 323–349.
Corsi, F., 2009. A simple approximate long-memory model of realized volatility. J. Financial Econometrics 7 (2), 174–196.
Corsi, F., Pirino, P., Reno, R., 2010. Threshold bipower variation and the impact of jumps on volatility forecasting. J. Econometrics 159 (2), 276–288.
Diebold, F., Mariano, R., 2002. Comparing predictive accuracy. J. Business & Economic Statistics 20 (1), 134–144.
Diebold, F.X., Strasser, G., 2013. On the correlation structure of microstructure noise: a financial economic approach. Rev. Econom. Stud. 80 (4), 1304–1337.
Driscoll, J.C., Kraay, A.C., 1998. Consistent covariance matrix estimation with spatially dependent panel data. Rev. Economics and Statistics 80 (4), 549–560.
French, K.R., Schwert, G.W., Stambaugh, R.F., 1987. Expected stock returns and volatility. Journal of Financial Economics 19 (1), 3–29.
Gatheral, J., Oomen, R.C.A., 2010. Zero-intelligence realized variance estimation. Finance Stoch. 14 (2), 249–283.
Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74 (6), 1545–1578.
Hansen, P., 2005. A test for superior predictive ability. J. Business & Economic Statistics 23 (4), 365–380.
Hansen, P., Lunde, A., 2005. A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?. J. Appl. Econometrics 20 (7), 873–889.
Hansen, P., Lunde, A., 2006a. Consistent ranking of volatility models. J. Econometrics 131 (1–2), 97–121.
Hansen, P.R., Lunde, A., 2006b. Realized variance and market microstructure noise. J. Business & Economic Statistics 24 (2), 127–161.
Hansen, P., Lunde, A., 2014. Estimating the persistence and the autocorrelation function of a time series that is measured with error. Econometric Theory 30 (1), 60–93.
Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79 (2), 453–497.
Hautsch, N., Podolskij, M., 2013. Preaveraging-based estimation of quadratic variation in the presence of noise and jumps: Theory, implementation, and empirical evidence. J. Business & Economic Statistics 31 (2), 165–183.
Hoechle, D., 2007. Robust standard errors for panel regressions with cross-sectional dependence. Stata J. 7 (3), 281.
Jacod, J., Li, Y., Mykland, P.A., Podolskij, M., Vetter, M., 2009. Microstructure noise in the continuous case: the pre-averaging approach. Stochastic Process. Appl. 119 (7), 2249–2276.
Jacod, J., Todorov, V., 2014. Efficient estimation of integrated volatility in presence of infinite variation jumps. Ann. Statist. 42 (3), 1029–1069.
Mancini, C., 2001. Disentangling the jumps of the diffusion in a geometric jumping Brownian motion. G. dell. Ital. degli Attuari 64, 19–47.
Mancini, C., 2009. Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. Scand. J. Stat. 36 (2), 270–296.
Martens, M., Van Dijk, D., 2007. Measuring volatility with the realized range. J. Econometrics 138 (1), 181–207.
Meddahi, N., 2003. ARMA representation of integrated and realized variances. Econom. J. 6 (2), 335–356.
Meddahi, N., Mykland, P., Shephard, N. (Eds.), 2011. Special issue on realised volatility. J. Econometrics 160 (1).

Oomen, R.C.A., 2006a. Comment. Journal of Business & Economic Statistics 24 (2), 195–202.
Oomen, R.C.A., 2006b. Properties of realized variance under alternative sampling schemes. J. Bus. Econom. Statist. 24, 219–237.
Parkinson, M., 1980. The extreme value method for estimating the variance of the rate of return. J. Business 53, 61–65.
Patton, A.J., 2011a. Data-based ranking of realised volatility estimators. J. Econometrics 161 (2), 284–303.
Patton, A.J., 2011b. Volatility forecast comparison using imperfect volatility proxies. J. Econometrics 160 (1), 246–256.
Patton, A.J., Sheppard, K., 2009a. Evaluating volatility and correlation forecasts. In: Andersen, T.G., Davis, R.A., Kreiss, J.-P., Mikosch, T. (Eds.), Handbook of Financial Time Series. Springer, Verlag.
Patton, A.J., Sheppard, K., 2009b. Optimal combinations of realised volatility estimators. Int. J. Forecast. 25 (2), 218–238.
Patton, A.J., Sheppard, K., 2013. Good volatility, bad volatility: Signed jumps and the persistence of volatility. Rev. Economics & Statistics forthcoming.
Podolskij, M., Vetter, M., 2009a. Bipower-type estimation in a noisy diffusion setting. Stochastic process. Appl. 119 (9), 2803–2831.
Podolskij, M., Vetter, M., 2009b. Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. Bernoulli 15 (3), 634–658.
Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. J. Amer. Statist. Assoc. 89 (428), 1303–1313.
Politis, D.N., White, H., 2004. Automatic block-length selection for the dependent bootstrap. Econometric Rev. 23 (1), 53–70.
Romano, J.P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73 (4), 1237–1282.
Vogelsang, T.J., 2012. Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. J. Econometrics 166 (2), 303–319.
White, H., 2000. A reality check for data snooping. Econometrica 68 (5), 1097–1126.
Zhang, L., 2006. Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. Bernoulli 12 (6), 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. J. Amer. Statist. Assoc. 100 (472), 1394–1411.
Zhou, B., 1996. High-frequency data and volatility in foreign-exchange rates. J. Bus. Econom. Statist. 14 (1), 45–52.