Realized Variance and Market Microstructure Noise

Peter R. Hansen[a]*, **Asger Lunde**[b]

[a] *Stanford University, Department of Economics, 579 Serra Mall, Stanford, CA 94305-6072, USA*

[b] *Aarhus School of Business, Department of Marketing, Informatics and Statistics, Fuglesangs Alle 4, 8210 Aarhus V, Denmark*

First version: January 2004. This version: September, 2004

Abstract

We examine a simple bias correction of the realized variance (*RV*) in the situation where observed prices are contaminated with market microstructure noise. The bias correction can greatly reduce the mean squared error of the *RV*, and we show that the bias corrected estimator can be utilized to uncover important characteristics of market microstructure noise. An empirical analysis of the 30 stocks that comprise the Dow Jones Industrial Average reveals that market microstructure noise is time-dependent and correlated with increments in the efficient price. These properties are found in both transaction and quotation data. Both characteristics have important implications for volatility estimation based on high-frequency data.

Keywords: Realized Variance; Realized Volatility; Integrated Variance; Market Microstructure Noise; Bias Correction; High-Frequency Data; Sampling Schemes.

JEL Classification: C10; C22; C80.

1. Introduction

The realized variance (*RV*) has become a popular empirical measure of volatility, and the *RV* yields a perfect estimate of volatility in the hypothetical situation where prices are observed in continuous time and without measurement error. This result suggests that the *RV*, which is a sum-of-squared returns, should be based on returns that are sampled at the highest possible frequency (tick-by-tick data). However, in practice this leads to a well-known bias problem due to market microstructure noise, see e.g. Zhou (1996), Andreou & Ghysels (2002), and Oomen (2002). The bias is particularly evident from *volatility signature plots* that were introduced by Andersen, Bollerslev, Diebold & Labys (2000), and the presence of noise has recently been documented with formal hypothesis tests by Awartani, Corradi & Distaso (2004). So there is a trade-off between bias and variance when choosing the sampling frequency, and this is the reason that returns are typically sampled at a moderate frequency, such as 5-minute sampling. An alternative way to handle the bias problem

*Corresponding author, email: peter.hansen@stanford.edu

is to use bias correction techniques, such as the filtering techniques that were used by Ebens (1999) and Andersen, Bollerslev, Diebold & Ebens (2001) (moving average filter) and Bollen & Inder (2002) (autoregressive filter). Other bias corrections were recently introduced by Zhang, Mykland & Aït-Sahalia (2002) (subsample approach) who consider time-independent noise, and by Hansen & Lunde (2003) (kernel-based approach) who allow for time-dependence in the noise process.

In this paper, we analyze an estimator that was introduced in this context by Zhou (1996).[1] We denote this estimator by RV_{AC_1}, because it utilize the first-order autocorrelation to bias-correct the RV. We make four contributions in this paper. First, we derive the properties of RV_{AC_1} in a slightly more general setting than did Zhou (1996), as we allow for non-constant volatility and non-Gaussian market microstructure noise. Further, we benchmark RV_{AC_1} to the standard measure of RV and show that the former is superior to the latter in terms of the mean squared error (MSE) criterion. When their respective 'optimal' sampling frequencies are employed, we find that the RV_{AC_1} may reduce the MSE by 50% or more, compared to the standard RV. Second, we evaluate the distortions from neglecting the market microstructure noise. Interestingly we find that the asymptotic results of Barndorff-Nielsen & Shephard (2002) provide reasonably accurate confidence intervals (for the integrated variance) at low sampling frequencies, such as 20-minute sampling. This finding is somewhat remarkably since Barndorff-Nielsen & Shephard (2002) derive their results in the absence of market microstructure effects. However, at five-minute sampling we find the "true" confidence interval to be 7%-30% larger than the confidence interval that is based on an assumed absence of noise. Third, we propose a simple test of the hypothesis of 'no time-dependence in the noise process'. The construction of this test exploits the bias corrected estimator's ability to uncover features of the latent noise process. This test is interesting because the absence of time-dependence in the noise process is commonly assumed when the effects of market microstructure noise are analyzed, see e.g. Corsi, Zumbach, Müller & Dacorogna (2001), Bandi & Russell (2003)[2], and Zhang, Mykland & Aït-Sahalia (2003). Fourth, we uncover some interesting features of market microstructure noise in an empirical analysis of returns for the equities that comprise the Dow Jones Industrial Average (DJIA). We find evidence that the noise process is both time-dependent and correlated with the returns of the efficient price. This finding is robust to the choice of sampling method (calendar-time or tick-time) and the type of price data (transaction prices or quotation prices). These two find-

[1] This estimator has previously been applied to daily return series by French, Schwert & Stambaugh (1987).

[2] A later version of the working paper by Bandi and Russell (which appeared after the first version of the present paper) relax the iid assumption and allow for a mild form of time-dependence. The new version of their paper also argues in favor of bias correcting the standard measure of RV.

ings have important implications for sampling intraday returns at ultra high frequencies, such as every few ticks or every few seconds. The dependence between the noise process and the efficient price process has important implications for some of the bias corrections that have been used in the literature.

The paper is organized as follows. We formulate assumptions and present theoretical results in Section 2 where we compare RV and RV_{AC_1} and evaluate the accuracy of 'no-noise' confidence intervals for the integrated variance. Section 3 contains our empirical analysis and presents a test for time-independence in market microstructure noise. Section 4 contains a summary and concluding remarks. All proofs are presented in the appendix.

2. Definitions and Theoretical Results

Let $\{p^*(t)\}$ be a latent log-price process in continuous time and let $\{p(t)\}$ be the observable log-prices process, such that the measurement error (noise) process is given by $u(t) \equiv p(t) - p^*(t)$. The noise process, u, may be due to market microstructure effects such as bid-ask bounces, but the discrepancy between p and p^* can also be induced by the technique that is used to construct $p(t)$. For example, p is often constructed artificially from observed trades and quotes using the *previous-tick* method or the *linear interpolation* method.[3]

We shall work under the following specification for the efficient price process, p^*.

Assumption 1 *The efficient price process satisfies $dp^*(t) = \sigma(t)dw(t)$, where $w(t)$ is a standard Brownian motion, $\sigma(t)$ is a time-varying (random) function that is independent of w, and $\sigma^2(t)$ is Lipschitz (almost surely).*

In our analysis we shall condition on the volatility path, $\{\sigma^2(t)\}$, because our object of interest is the *integrated variance*,

$$IV \equiv \int_a^b \sigma^2(t)dt.$$

So we can treat $\{\sigma^2(t)\}$ as deterministic even though the volatility path is considered to be random. The Lipschitz condition is a smoothness condition that requires $|\sigma^2(t) - \sigma^2(t+\delta)| < \epsilon\delta$ for some ϵ and all t and δ (with probability one).

Next we formulate assumptions about the noise process, where "$\perp$" refers to stochastic independence.

[3] The former was proposed by Wasserfallen & Zimmermann (1985) and the latter was used by Andersen & Bollerslev (1997). For a discussion of the two, see Dacorogna, Gencay, Müller, Olsen & Pictet (2001, sec. 3.2.1).

Assumption 2 *The noise process satisfies:*

(i) $p^* \perp u$; $u(s) \perp u(t)$ *for all* $s \neq t$; *and* $E[u(t)] = 0$ *for all* t;

(ii) $\omega^2 \equiv E|u(t)|^2 < \infty$ *for all* t;

(iii) $\mu_4 \equiv E|u(t)|^4 < \infty$ *for all* t.

Assumption 2.i will be maintained throughout our analysis, whereas (ii) and (iii) will only be made when necessary. To simplify some of our subsequent expressions we define the "excess kurtosis ratio", $\kappa \equiv \mu_4/(3\omega^4)$. Assumption 2 is clearly satisfied if u is a Gaussian 'white noise' process, $u(t) \sim N(0, \omega^2)$, in which case $\kappa = 1$.

The existence of a noise process, u, that satisfies Assumption 2, follows directly from Kolmogorov's existence theorem, see Billingsley (1995, chapter 7). It is worthwhile to note that 'white noise processes in continuous time' are very erratic processes. In fact, the quadratic variation of an white noise process is unbounded (as is the *r-tic variation* for any other integer). So the 'realized variance' of an white process diverges to infinity as the sampling frequency is increased. This is in stark contrast to the situation for Brownian motion type processes that have finite r-tic variation for $r \geq 2$, see Barndorff-Nielsen & Shephard (2003).

2.1. *Sampling Scheme*

We partition the interval $[a, b]$ into m subintervals, and the number of subintervals, m, plays a central role in our analysis. For example we shall derive asymptotic distributions of quantities, as $m \to \infty$, and discuss optimal choices for m. For a fixed m the ith subinterval is given by $[t_{i-1,m}, t_{i,m}]$, where $a = t_{0,m} < t_{1,m} < \cdots < t_{m,m} = b$. The length of the ith subinterval is given by $\delta_{i,m} \equiv t_{i,m} - t_{i-1,m}$ and we assume that $\sup_{i=1,\ldots,m} \delta_{i,m} = O(\frac{1}{m})$, such that the length of each subinterval shrinks to zero as m increases. The *intraday returns* are now defined by,

$$y^*_{i,m} \equiv p^*(t_{i,m}) - p^*(t_{i-1,m}), \qquad i = 1, \ldots, m,$$

and the increments in p and u are defined similarly and denoted by $y_{i,m}$ and $e_{i,m}$, respectively. Note that the *observed intraday returns* decomposes into $y_{i,m} = y^*_{i,m} + e_{i,m}$. Next we define the integrated variance over each of the subintervals,

$$\sigma^2_{i,m} \equiv \int_{t_{i-1,m}}^{t_{i,m}} \sigma^2(s) ds, \qquad i = 1, \ldots, m,$$

and note that $\text{var}(y^*_{i,m}) = E(y^{*2}_{i,m}) = \sigma^2_{i,m}$.

The *realized variance* of p^* is defined by $RV_*^{(m)} \equiv \sum_{i=1}^m y_{i,m}^{*2}$, and it follows that $RV_*^{(m)}$ is consistent for the *IV*, as $m \to \infty$, see e.g. Meddahi (2002). An asymptotic distribution theory of realized variance (in relation to integrated variance) is established in Barndorff-Nielsen & Shephard (2002). While $RV_*^{(m)}$ is an ideal estimator – it is not a feasible estimator – because p^* is latent. The realized variance of p, which is given by $RV^{(m)} \equiv \sum_{i=1}^m y_{i,m}^2$, is observable but suffers from a well-known bias problem and is inconsistent for the *IV*.

The literature has been concerned with different forms of sampling schemes. The special case where $t_{i,m}$, $i = 1, \ldots, m$ are equidistant in time, i.e. $\delta_{i,m} = (b-a)/m$ for all i, is referred to as *calendar time sampling* (CTS). The widely used exchange rates data from Olsen and Associates, see Müller, Dacorogna, Olsen, Pictet, Schwarz & Morgenegg (1990) are equidistant in time, and five-minute sampling ($\delta_{i,m} = 5\,\text{min}$) has commonly been used in this context. The case where the sampling times, $t_{0,m}, \ldots, t_{m,m}$, are such that $\sigma_{i,m}^2 = IV/m$ for all $i = 1, \ldots, m$, is referred to as *business time sampling* (BTS), see Oomen (2004*b*). Whereas, the case where $t_{i,m}$ refers to the time of a transaction/quotation, will be referred to as *tick time sampling* (TTS). An example of TTS is when $t_{i,m}$, $i = 1, \ldots, n$, are chosen to be the time of every fifth transaction, say. While $t_{i,m}$, $i = 0, \ldots, m$, are observable under CTS and TTS, they are latent under BTS, because the sampling times are defined from the unobserved volatility path. Yet empirical results by Andersen & Bollerslev (1997) and Curci & Corsi (2004) suggest that BTS can be approximated by TTS. This feature is nicely captured in the framework of Oomen (2004*b*), where the (random) tick times are generated with an intensity that is directly related to a quantity that corresponds to $\sigma^2(t)$ in the present context. Under CTS we will sometimes write $RV^{(x\,\text{sec})}$, where x seconds is the period in time spanned by each of the intraday returns (i.e. $\delta_{i,m} = x$ seconds). Similarly, we write $RV^{(y\,\text{ticks})}$ under TTS when each intraday return spans y ticks (transactions or quotations).

The bias-variance properties of the $RV^{(m)}$ have been analyzed by Bandi & Russell (2003) and Zhang et al. (2003) under an white noise assumption. The following lemma summarizes these results in our notation, where we use '$\xrightarrow{d}$' to denote convergence in distribution.

Lemma 1 *Conditional on* $\{\sigma^2(s)\}$ *and given Assumptions 1 and 2.i-ii it holds that* $E(RV^{(m)}) = IV + 2m\omega^2$; *if Assumption 2.iii also holds, then*

$$\text{var}(RV^{(m)}) = \kappa 12\omega^4 m + 8\omega^2 \sum_{i=1}^m \sigma_{i,m}^2 - (6\kappa - 2)\omega^4 + 2\sum_{i=1}^m \sigma_{i,m}^4,$$

and

$$\frac{RV^{(m)} - 2m\omega^2}{\sqrt{\kappa 12\omega^4 m}} = \sqrt{\frac{m}{3\kappa}}\left(\frac{RV^{(m)}}{2m\omega^2} - 1\right) \xrightarrow{d} N(0,1), \qquad \text{as } m \to \infty.$$

In the absence of market microstructure noise and under CTS ($\omega^2 = 0$ and $\delta_{i,m} = (b-a)/m$), we obtain a result of Barndorff-Nielsen & Shephard (2002), that

$$\text{var}(RV^{(m)}) = 2\sum_{i=1}^{m} \sigma_{i,m}^4 = \tfrac{2}{m}\int_a^b \sigma^4(s)ds + o(\tfrac{1}{m}),$$

where $\int_a^b \sigma^4(s)ds$ is known as the *integrated quarticity* that was introduced by Barndorff-Nielsen & Shephard (2002).

Next, we consider the estimator of Zhou (1996) that is given by

$$RV_{\text{AC}_1}^{(m)} \equiv \sum_{i=1}^{m} y_{i,m}^2 + \sum_{i=1}^{m} y_{i,m}y_{i-1,m} + \sum_{i=1}^{m} y_{i,m}y_{i+1,m}. \tag{1}$$

This estimator incorporates the empirical first-order autocorrelation, which amounts to a bias correction that 'works' the same way that robust covariance estimators, such as that of Newey & West (1987), achieve their consistency. Note that (1) involves $y_{0,m}$ and $y_{m+1,m}$ that are intraday returns outside the interval $[a, b]$. We choose this formulation because it simplifies several expressions (by avoiding $m/(m-1)$ terms). If these two intraday returns are unavailable, one could simply use the estimator $\sum_{i=2}^{m-1} y_{i,m}^2 + \sum_{i=2}^{m} y_{i,m}y_{i-1,m} + \sum_{i=1}^{m-1} y_{i,m}y_{i+1,m}$ that estimates $\int_{a+\delta_{1,m}}^{b-\delta_{m,m}} \sigma^2(s)ds = IV + O(\frac{1}{m})$.

In the following lemma we establish results for $RV_{\text{AC}_1}^{(m)}$ that are similar to those for $RV^{(m)}$ in Lemma 1.

Lemma 2 *Conditional on $\{\sigma^2(s)\}$ and given Assumptions 1 and 2.i it holds that $E(RV_{\text{AC}_1}^{(m)}) = IV$; if Assumption 2.ii also holds then*

$$\text{var}(RV_{\text{AC}_1}^{(m)}) = 8\omega^4 m + 8\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2 - 6\omega^4 + 6\sum_{i=1}^{m} \sigma_{i,m}^4 + r_m,$$

where $r_m = O(m^{-2})$ under CTS and BTS, and

$$\frac{RV_{\text{AC}_1}^{(m)} - IV}{\sqrt{8\omega^4 m}} \xrightarrow{d} N(0,1), \qquad \text{as} \quad m \to \infty.$$

An important result of Lemma 2 is that $RV_{\text{AC}_1}^{(m)}$ is unbiased for the *IV* at any sampling frequency, m, and that this result is established under relatively weak assumptions about the noise process. Also note that Lemma 2 requires weaker assumptions that those needed for $RV^{(m)}$ in Lemma 1. This is achieved because the expression for $RV_{\text{AC}_1}^{(m)}$ can be rewritten such that it does not involve squared noise terms, $[u_{i,m}]^2$, $i = 1, \ldots, m$, as do $RV^{(m)}$, where $u_{i,m} \equiv u(t_{i,m})$. A somewhat remarkable result of Lemma 2 is that the bias corrected estimator, $RV_{\text{AC}_1}^{(m)}$, has a smaller asymptotic variance than

the unadjusted estimator, $RV^{(m)}$. Usually a bias correction is accompanied by a larger asymptotic variance. Also note that the asymptotic results of Lemma 2 is more useful than that of Lemma 1 (in terms of estimating IV), because the result of Lemma 1 does not involve the object of interest, IV, but only shed light on aspects of the noise process. However, it is also important to note that the asymptotic result of Lemma 2 does not suggest that $RV^{(m)}_{\text{AC}_1}$ should be based on intraday returns that are sampled at the highest possible frequency, since the asymptotic variance is increasing in m! In other words: While $RV^{(m)}_{\text{AC}_1}$ is centered about the object of interest, IV, it is unlikely to be close to IV as $m \to \infty$.

With CTS ($\delta_{i,m} = (b-a)/m$) our expression for the variance is approximately given by

$$\text{var}[RV^{(m)}_{\text{AC}_1}] \approx 8\omega^4 m + 8\omega^2 \int_a^b \sigma^2(s) - 6\omega^4 + 6\int_a^b \sigma^4(s)ds\frac{1}{m}.$$

This shows that the variance of $RV^{(m)}_{\text{AC}_1}$ is about three times larger than that of $RV^{(m)}$ in the absence of market microstructure noise ($\omega^2 = 0$), for a given m. This shortcoming can be compensated for by using a sampling frequency for $RV^{(m_1)}_{\text{AC}_1}$ that is three times higher than that of $RV^{(m_0)}$ (i.e. $m_1 = 3m_0$). Our empirical results suggest that $RV^{(m)}_{\text{AC}_1}$ should be based on intraday returns that are sampled ten times more frequently than those for $RV^{(m_1)}$, such that this component of the total variance is reduced.

Next, we compare $RV^{(m)}_{\text{AC}_1}$ to $RV^{(m)}$ in terms of their mean square error (MSE) and their respective optimal sampling frequencies for a special case that illustrates key features of the two estimators.

Corollary 3 *Define $\lambda \equiv \omega^2/IV$, suppose that $\kappa = 1$, and let $t_{0,m}, \ldots, t_{m,m}$ be such that $\sigma^2_{i,m} = IV/m$ (BTS). The mean squared errors are given by*[4]

$$\text{MSE}(RV^{(m)}) = IV^2[4\lambda^2 m^2 + 12\lambda^2 m + 8\lambda - 4\lambda^2 + \tfrac{2}{m} \quad], \tag{2}$$

$$\text{MSE}(RV^{(m)}_{\text{AC}_1}) = IV^2[\quad\quad 8\lambda^2 m + 8\lambda - 6\lambda^2 + \tfrac{6}{m} - \tfrac{2}{m^2}]. \tag{3}$$

The optimal sampling frequencies for $RV^{(m)}$ and $RV^{(m)}_{\text{AC}_1}$ are given implicitly as the real (positive) solutions to $4\lambda^2 m^3 + 6\lambda^2 m^2 - 1 = 0$ and $4\lambda^2 m^3 - 3m + 2 = 0$, respectively.

We denote the optimal sampling frequencies for $RV^{(m)}$ and $RV^{(m)}_{\text{AC}_1}$ by m_0^* and m_1^*, respectively, and these are approximately given by, $m_0^* \approx (2\lambda)^{-2/3}$ and $m_1^* \approx \sqrt{3}(2\lambda)^{-1}$. In our empirical analysis we find λ^{-1} to be larger than $1,000$ for most equities, such that $m_1^*/m_0^* \approx 3^{1/2}2^{-1/3}(\lambda^{-1})^{1/3} \geq 10$,

[4]Equation (3) was derived in Zhou (1996) under the assumption that the noise process is Gaussian. However, his expression contains a minor error as "$-6\lambda^2$" is incorrectly written as "$-4\lambda^2$" (using our notation).

and this shows that m_1^* is several times larger than m_0^* in practice. In other words, the optimal $RV_{\mathrm{AC}_1}^{(m)}$ requires a more frequent sampling than does the 'optimal' RV. This is quite intuitive, because $RV_{\mathrm{AC}_1}^{(m)}$ can utilize more information in the data without being affected by a severe bias. Naturally, when using TTS, the number of intraday returns, m, cannot exceed the total number of transactions/quotations, so in practice it might not be possible to sample as frequently as prescribed by m_1^*.

Although Corollary 3 is based on somewhat restrictive assumptions (such as $\sigma_{i,m}^2 = IV/m$) it nevertheless captures the salient features of this problem, and the MSE properties (in relation to the noise process) are characterized by a single parameter, λ. So Corollary 3 provide a very attractive framework for comparing $RV^{(m)}$ and $RV_{\mathrm{AC}_1}^{(m)}$, and for evaluating how sensitive these estimators are to market microstructure noise.

[FIGURE 1 ABOUT HERE]

From Corollary 3 we note that the root mean squared errors (RMSEs) of $RV^{(m)}$ and $RV_{\mathrm{AC}_1}^{(m)}$ are proportional to the IV and given by $\gamma_0(\lambda, m)IV$ and $\gamma_1(\lambda, m)IV$, respectively, where

$$\gamma_0(\lambda, m) \equiv \sqrt{4\lambda^2 m^2 + 12\lambda^2 m + 8\lambda - 4\lambda^2 + \tfrac{2}{m}},$$

$$\gamma_1(\lambda, m) \equiv \sqrt{8\lambda^2 m + 8\lambda - 6\lambda^2 + \tfrac{6}{m} - \tfrac{2}{m^2}}.$$

In Figure 1 we have plotted $\gamma_0(\lambda, m)$ and $\gamma_1(\lambda, m)$ using two empirical estimates of λ. The estimates are based on high-frequency stock returns of AA:Alcoa Inc. (upper panels) and Microsoft:MSFT (lower panels). The details regarding the estimation of λ is deferred to the empirical section of this paper. The left panels present $\gamma_0(\hat{\lambda}, m)$ and $\gamma_1(\hat{\lambda}, m)$, where the the x-axis is $\delta_{i,m} = (b-a)/m$ in units of seconds. For both equities, we note that the $RV_{\mathrm{AC}_1}^{(m)}$ dominate the $RV^{(m)}$ except at the very lowest frequencies. The minimum of $\gamma_0(\hat{\lambda}, m)$ and $\gamma_1(\hat{\lambda}, m)$ identify their respective optimal sampling frequencies, m_0^* and m_1^*. For the AA returns we find the optimal sampling frequencies to be $m_{0,AA}^* = 77$ and $m_{1,AA}^* = 1190$ and the theoretical reduction of the MSE is 64.5%. The curvature of $\gamma_0(\hat{\lambda}, m)$ and $\gamma_1(\hat{\lambda}, m)$ in the neighborhood of m_0^* and m_1^*, respectively, show that $RV_{\mathrm{AC}_1}^{(m)}$ is less sensitive to the choice of m than is $RV^{(m)}$.

The right panels of Figure 1 display the relative MSE of $RV_{\mathrm{AC}_1}^{(m)}$ to that of (the optimal) $RV^{(m_0^*)}$ and the relative MSE of $RV^{(m)}$ to that of (the optimal) $RV_{\mathrm{AC}_1}^{(m_1^*)}$. These panels show that the $RV_{\mathrm{AC}_1}^{(m)}$ continue to dominate the 'optimal' $RV^{(m_0^*)}$ for a wide ranges of frequencies, and not just in a small

neighborhood of the optimal value, m_1^*. This robustness of RV_{AC_1} is quite useful in practice where λ and (hence) m_1^* are not known with certainty, because the result shows that a reasonably precise estimate of λ (and hence m_1^*) will lead to an RV_{AC_1} that dominates the 'optimal' RV.

[FIGURE 2 ABOUT HERE]

A second very interesting aspect that can be analyzed from the results of Corollary 3, is the accuracy of theoretical results that are derived under the assumption that $\lambda = 0$ (no market microstructure noise). For example, the accuracy of a confidence interval for IV, which is based on the asymptotic results of Barndorff-Nielsen & Shephard (2002), will depend on λ and m, and the expressions of Corollary 3 provide a simple way to quantify the accuracy of such confidence intervals. Figure 2 provide valuable information about this question. The left panels of Figure 2 present the RMSEs of $RV^{(m)}$ and $RV_{\text{AC}_1}^{(m)}$, using both an estimate $\hat{\lambda} > 0$ (the case with noise) and $\lambda = 0$ (the case without noise). For small values of m we see that $\gamma_0(\hat{\lambda}, m) \approx \gamma_0(0, m)$ and $\gamma_1(\hat{\lambda}, m) \approx \gamma_1(0, m)$, whereas the effects of market microstructure noise are pronounced at the higher sampling frequencies. The right panels of Figure 2 quantify the discrepancy between the two 'types' of confidence intervals (about $RV^{(m)}$) as a function of the sampling frequency. These plots present $100[\gamma_0(\hat{\lambda}, m) - \gamma_0(0, m)]/\gamma_0(0, m)$ and $100[\gamma_1(\hat{\lambda}, m) - \gamma_1(0, m)]/\gamma_1(0, m)$ as a function of m. So the former reveals the percentage widening of a confidence intervals for IV (about $RV^{(m)}$) due to market microstructure noise, and the second line shows the corresponding widening of the confidence interval that is constructed about $RV_{\text{AC}_1}^{(m)}$. The vertical lines in the right panels mark the sampling frequency that corresponds to five-minute sampling under CTS, and these show that the 'actual' confidence interval (based on $RV^{(m)}$) is 31.47% larger than the 'no-noise' confidence interval for AA, whereas the enlargement is 7.57% for MSFT. At 20-minute sampling the discrepancy is less than a couple of percent, so in this case the sized distortion from being oblivious to market microstructure noise is quite small. The corresponding increases in the RMSE of $RV_{\text{AC}_1}^{(m)}$ are 3.88% and 1.44%, respectively. So a 'no-noise' confidence interval (about $RV_{\text{AC}_1}^{(m)}$) could, in principle, be used as a reasonable approximation at moderate sampling frequencies (if a better alternative is not available).

3. Empirical Analysis

We analyze stock returns for the 30 equities of the Dow Jones Industrial Average (DJIA). The sample period spans three years, from January 3, 2000 to December 31, 2002, which delivers a total of 750

trading days. The data are transaction prices and quotations from the two exchanges NYSE and NASDAQ, and all data were extracted from the Trade and Quote (TAQ) database.

The raw data were filtered for outliers and we discarded transactions outside the period from 9:30am to 4:00pm, and days with less than five hours of trading were removed from the sample, which reduced the sample to the number of days reported in the last column of Table 1. The average number of transactions/quotations per day are labelled (All) in Table 1. Our analysis of quotation data is based on mid-quotes (average of bid and ask prices). When sampling in calendar time we used the previous-tick method to construct intraday returns, and when sampling in tick-time we remove price observations that did not result in a 'price change'.[5] The average number of 'new' transaction/quotation prices per day is labelled ($\#\Delta p_{t_i} \neq 0$) in Table 1. The *RV*s are calculated for the hours that the market is open, approximately 390 minutes per day (6.5 hours for most days), and we denote the *RV*s on day t, by $RV_t^{(m)}$ and $RV_{\mathrm{AC}_1,t}^{(m)}$, $t = 1, \ldots, n$. We present result for all 30 equities in our tables, whereas the figures present results for two equities: Alcoa Inc. (AA) and Microsoft (MSFT), that represent equities of the DJIA with low and high trading activities, respectively. The corresponding figures for the other 28 DJIA equities are available upon request.

[TABLE 1 ABOUT HERE: Data Description]

3.1. *Estimation of Market Microstructure Noise Parameters*

From Lemmas 1 and 2 it follows that $2m\omega^2 = E[RV^{(m)} - RV_{\mathrm{AC}_1}^{(\tilde{m})}]$, where $\tilde{m}$ is arbitrary. Under the assumption that ω^2 is constant across days, it can be estimated by

$$\hat{\omega}^2 \equiv n^{-1} \sum_{t=1}^{n} \frac{1}{2m_t} (RV_t^{(m_t)} - RV_{\mathrm{AC}_1,t}^{(\tilde{m}_t)}), \tag{4}$$

where m_t and $\tilde{m}_t$ need not be constant across days. It the special case where the same frequencies are used for all days, (4) simplifies to $\hat{\omega}^2 = \frac{1}{2m}(\overline{RV}^{(m)} - \overline{RV}_{\mathrm{AC}_1}^{(\tilde{m})})$, where $\overline{RV}^{(m)} \equiv n^{-1} \sum_{t=1}^{n} RV_t^{(m)}$ and $\overline{RV}_{\mathrm{AC}_1}^{(\tilde{m})} \equiv n^{-1} \sum_{t=1}^{n} RV_{\mathrm{AC}_1,t}^{(\tilde{m})}$.

In order to obtain a precise estimate of ω^2 it is important that $RV_{\mathrm{AC}_1,t}^{(\tilde{m}_t)}$ is indeed unbiased for *IV*. For this reason, we calculate $RV_{\mathrm{AC}_1,t}$ using tick-time sampling, where each of the intraday returns spans 60 transactions (or quotations for the mid-quote data). Thus $\tilde{m}_t$ will (approximately) equal

[5] The same price is often observed in several consecutive transaction/quotations, because a large trade is divided into smaller transactions (transaction data) and because the market maker issues a new quote with a different 'depth' while the bid and ask prices are unchanged (quotation data). This censoring does not affect *RV* but is important for RV_{AC_1}, because it effectively removes all the zero-intraday returns which influence the autocovariances of intraday returns.

the number of transactions (or quotations) on day t divided by 60. The higher the frequency that is used for $RV^{(m)}$, the more dominating is the bias term, so we choose a much higher frequency in this case. Specifically we use one-tick sampling, such that our results are based on $RV_t^{(m_t)} = RV_t^{(1\text{ ticks})}$ and $RV_{\text{AC}_1,t}^{(\tilde{m}_t)} = RV_{\text{AC}_1,t}^{(60\text{ ticks})}$.

Under our assumptions it follows that $\text{plim}_{m\to\infty}\frac{1}{2m}(RV^{(m)} - RV_{\text{AC}_1}^{(\tilde{m})}) = \text{plim}_{m\to\infty}\frac{1}{2m}RV^{(m)}$, and $\frac{1}{2m}RV^{(m)}$ is used (as a consistent estimator of ω^2) in Zhang et al. (2003) to bias correct their subsample estimator. In our empirical analysis we do not find $RV_{\text{AC}_1}^{(\tilde{m})}$ to be negligible, even when sampling at the highest possible frequency, 1-tick sampling. We observed a substantial difference between the two estimates, $\frac{1}{2m}(RV^{(m)} - RV_{\text{AC}_1}^{(\tilde{m})})$ and $\frac{1}{2m}RV^{(m)}$, so even though both are consistent for ω^2, the latter is quite biased for at the sampling frequency, m, that one can use in practice. So we argue that it is important to incorporate RV_{AC_1}, or some other unbiased estimator of IV, whenever ω^2 is estimated in this way.

Since $RV_{\text{AC}_1}^{(\tilde{m})}$ is unbiased for IV, it follows that $\overline{IV} \equiv n^{-1}\sum_{t=1}^{n} RV_{\text{AC}_1,t}^{(\tilde{m}_t)}$ estimates the average daily IV over the sample $t = 1, \ldots, n$. So we define

$$\hat{\lambda} \equiv \hat{\omega}^2 \,/\, \overline{IV}\,,$$

which is an estimator of 'average noise' divided by 'average integrated variance'. If both ω^2 and IV are constant across days, such that λ is the same for all days, then $\hat{\lambda}$ is consistent for λ under mild regularity conditions. In practice, λ is unlikely to be constant across days, so $\hat{\lambda}$ should be viewed as a reasonable approximation of λ, for a typical trading day. Fortunately, Figure 1 showed that $RV_{\text{AC}_1}^{(m)}$ is relatively insensitive to small deviations from m_1^*, such that an $\hat{m}_1^*$ based on a reasonable approximation for λ, leads to a more accurate estimator than $RV^{(m)}$ (for any m).

[TABLE 2 ABOUT HERE]

Table 2 contains empirical results for all 30 equities using both transaction and quotation data. The are several interesting observations to be made from Table 2. For the transaction data we note that $\hat{\lambda}$ is typically found to be smaller than 0.1%, and the theoretical reduction of the MSE, $100[\gamma_0^2(\hat{\lambda}, m_0^*) - \gamma_1^2(\hat{\lambda}, m_1^*)]/\gamma_0^2(\hat{\lambda}, m_0^*)$, is always found to be 50% or more. For example for Alcoa Inc. we find that $\hat{\omega}_{AA}^2 = 0.4217\%$ and $\hat{\lambda}_{AA} = 0.4217\%/5.797 = 0.0727\%$, that leads to the optimal sampling frequencies: $m_0^* = 77$ and $m_1^* = 1190$. For a typical trading day that is 6.5 hours long this corresponds to intraday returns that (on average) spans 5 minutes and 20 seconds, respectively. Bandi & Russell (2003) and Oomen (2004*b*, 2004*a*) report 'optimal' sampling frequencies for $RV^{(m)}$

that are similar to our estimates of m_0^*. By plugging these numbers into the formulae of Corollary 3 we find the reduction of the MSE (from using $RV_{\mathrm{AC}_1}^{(m_1^*)}$ rather than $RV^{(m_0^*)}$) to be 64.5%, which suggests that $RV_{\mathrm{AC}_1}^{(m_1^*)}$ is about three time more efficient than $RV^{(m_0^*)}$, provided that Assumptions 1 and 2 hold.

The noise-to-signal ratio, λ, is likely to differ across days, in which case the optimal sampling frequencies, m_0^* and m_1^*, will also differ across days. So our estimates above should be viewed as approximations for 'daily average values', in the sense that $m_0 = 148$ and $m_1 = 3139$ appear to be a sensible sampling frequencies to use for the MSFT transaction data.

For the quotation data all our estimates of ω^2 are negative, a phenomenon that also occurs for transaction data for three equities. A negative estimate occurs when the sample average of $RV^{(1\,\text{tick})}$ is smaller than the sample average of $RV_{\mathrm{AC}_1}^{(60\,\text{ticks})}$. This is obviously in conflict with the results of Lemmas 1 and 2 that dictate the population difference, $E[RV^{(1\,\text{tick})} - RV_{\mathrm{AC}_1}^{(60\,\text{ticks})}]$, to be positive. The expected difference is $2\omega^2$ times a number that is proportional to the average number of transactions/quotations per day. One explanation for observing $\hat{\omega}^2 < 0$ is that $\omega^2 \simeq 0$, such that the 'wrong' sign simply occurs by chance. However, this is highly improbable because all estimates of ω^2 (and not just about half of them) are found to be negative for the quotation data. Our subsequent analysis will also show that the negative estimates are caused by a violation of Assumption 2.

3.2. *A Hausman-Type Test for Time-Independence of the Noise Process*

From Lemma 2 we know that $RV_{\mathrm{AC}_1}^{(m)}$ is unbiased for IV at any frequency under Assumption 2.i. This followed from the fact that the innovations of the noise process, $e_{i,m}$, have a non-zero first-order autocovariance whereas higher-order autocovariances are all zero. This property is inherited by the observed intraday returns, $y_{i,m}$, given our assumptions about the efficient price process. The estimator, $RV_{\mathrm{AC}_1}^{(m)}$, is unbiased because it properly corrects for the first-order autocorrelation in $y_{i,m}$. However, $RV_{\mathrm{AC}_1}^{(m)}$ may be biased if higher-order autocovariances are non-zero, which would be the case (for large m) if the noise process, $u(t)$, was dependent across time (a violation of Assumption 2.i.)

[FIGURE 3 ABOUT HERE]

An graphical way to investigate the bias properties of RV-type estimators is through the so-called volatility signature plots of Andersen et al. (2000). In Figure 3 we present volatility signature

plots for AA and MSFT using both CTS and TTS, and based on both transaction data and quotation data. The upper four panels of Figure 3 are based on CTS and these reveal a pronounced bias when $RV^{(m)}_{\mathrm{AC}_1}$ is based on intraday returns that are sampled more frequently than every 30 seconds. The main explanation for this is that CTS will sample the same price multiple times when m is large, which induces (artificial) autocorrelation in intraday returns. Thus, when intraday returns are based on CTS, it is necessary to incorporate higher-order autocovariances of $y_{i,m}$, when m becomes large, see Hansen & Lunde (2004). The lower four plots are signature plots for TTS and these also reveal a pronounced bias in $RV^{(x \text{ ticks})}_{\mathrm{AC}_1}$ at the highest frequencies. Yet, a comparison of the CTS and TTS signature plots suggests that TTS dominates CTS, as it allows from a more frequent sampling of intraday returns. This observation is in line with the results of Oomen (2004*b*, 2004*a*), who showed this to be the case in a jump-process framework.

Another very important result of Figure 3, is that the volatility signature plots drops (rather than increases) as the sampling frequency increases (as $\delta_{i,m} \to 0$). This holds for both CTS and TTS, and both $RV^{(m)}$ and $RV^{(m)}_{\mathrm{AC}_1}$. This phenomenon was first documented by Hansen & Lunde (2004), and their theoretical results show that the negative bias of $RV^{(m)}$ cannot be explained by time-dependence in the noise process alone. As easy way to see this is to note that the quadratic variation of p is the sum of the quadratic variations of p^* and u whenever $u \perp p^*$. So the bias of (a nontrivial) market microstructure noise is always positive if $u \perp p^*$ holds. However, the volatility signature plots reveal a negative bias, and this strongly suggest that the innovations in the noise process, $e_{i,m}$, are negatively correlated with the true returns, $y^*_{i,m}$, see Hansen & Lunde (2004, theorem 2).

Because the unbiasness of $RV^{(m)}_{\mathrm{AC}_1}$ crucially relies on Assumption 2.i, we consider tests of the hypothesis,

$$H_0 : u \perp p^* \quad \text{and} \quad u(s) \perp u(t) \quad \text{for all } s \neq t.$$

As is almost evident from Figure 3, there is an overwhelming empirical evidence against H_0. So there is not need for a very powerful (or an optimal) test in order to conclude that H_0 is false. Instead we apply significance tests that can identify the reasons that H_0 does not hold. For this purpose we employ simple t-tests that have power against particular alternatives, and this will help us uncover the properties of market microstructure noise, and allow us to characterize the specific violation of Assumption 2.i that is found in these data.

For the construction of the tests we observe that $RV^{(m)}_{\mathrm{AC}_1}$ is unbiased for the IV for any m, under the null hypothesis, such that $RV^{(m)}_{\mathrm{AC}_1} = IV + error^{(m)}$ where $E(error^{(m)}) = 0$. It follows that

$$d_t \equiv RV^{(m)}_{\mathrm{AC}_1,t} - RV^{(\tilde{m})}_{\mathrm{AC}_1,t} = error^{(m)}_t - error^{(\tilde{m})}_t, \qquad m \neq \tilde{m},$$

is the difference between two measurement errors, each having expected value zero, such that $E[d_t] = 0$. The variance of the sample average, $\bar{d} \equiv n^{-1}\sum_{t=1}^{n} d_t$ (normalized by $\sqrt{n}$) is consistently estimated by $\frac{1}{n-1}\sum_{t=1}^{n}(d_t - \bar{d})^2$, because the measurements errors (and hence d_t) can be assumed to be uncorrelated across days. Thus a simple test of H_0, can be based on the following t-statistic,

$$t_{(m,\tilde{m})} \equiv \sqrt{n}\bar{d} \Big/ \sqrt{\tfrac{1}{n-1}\textstyle\sum_{t=1}^{n}(d_t - \bar{d})^2}\ ,$$

and under the null hypothesis we have that $t_{(m,\tilde{m})} \xrightarrow{d} N(0,1)$, as $n \to \infty$, under standard regularity conditions that are quite plausible to hold in this context.

The power of the test that compares $t_{(m,\tilde{m})}$ to a critical value of a standard Gaussian distribution, will depend on m and $\tilde{m}$. This can be seen from the decomposition of $RV^{(m)}_{\mathrm{AC}_1}$ into:

$$RV^{(m)}_{\mathrm{AC}_1} = \sum_{i=1}^{m} y_{i,m}(y_{i-1,m} + y_{i,m} + y_{i+1,m}) = \sum_{i=1}^{m}(\zeta^{yy}_{i,m} + \zeta^{uu}_{i,m} + \zeta^{yu}_{i,m}),$$

where

$$\begin{aligned}
\zeta^{yy}_{i,m} &= y^*_{i,m}(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m}) \\
\zeta^{uu}_{i,m} &= (u_{i,m} - u_{i-1,m})(u_{i+1,m} - u_{i-2,m}) \\
\zeta^{yu}_{i,m} &= y^*_{i,m}(u_{i+1,m} - u_{i-2,m}) + (u_{i,m} - u_{i-1,m})(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m}).
\end{aligned}$$

The expected value of the first term is $\sigma^2_{i,m}$, whereas the last two terms both have expected value zero under the null hypothesis. The decomposition also shows that autocorrelation in $u(t)$ may cause $E(\zeta^{uu}_{i,m}) \neq 0$ and correlation between increments in the efficient price and the noise process, can cause the expected value of the third term to be non-zero, $E(\zeta^{yu}_{i,m}) \neq 0$. So we see that the test that is based on $t_{(m,\tilde{m})}$ will have power against alternatives for which the bias of $RV^{(m)}_{\mathrm{AC}_1}$ differs from that of $RV^{(\tilde{m})}_{\mathrm{AC}_1}$, So the test will not have power against alternatives where both estimators are unbiased. This will occur if the time-dependence in u and the correlation between u and $y^*_{i,m}$ is short lived. For example, if the time-dependence in the noise process is less than one minute, then it is easy to verify that $RV^{(m)}_{\mathrm{AC}_1}$ is unbiased if m corresponds to intraday returns that each spans a period of time that is one minute or more. A similar example can be given for the case with TTS.

[TABLE 3 ABOUT HERE:]

[TABLE 4 ABOUT HERE:]

Tables 3 and 4 contain $t_{(m,\tilde{m})}$-statistics for pairs of $(m, \tilde{m})$. We chose $\tilde{m} = 60$ in most cases, because $RV_{\mathrm{AC}_1}^{(60\,\mathrm{ticks})}$ is unlikely to be biased, such that a rejection of H_0 based on $t_{(m,\tilde{m})}$ must be attributed to $RV_{\mathrm{AC}_1}^{(m)}$ being biased (for $m < \tilde{m}$). Thus our test is a Hausman-type test, because the sample averages of $RV_{\mathrm{AC}_1,t}^{(x\,\mathrm{ticks})}$ and $RV_{\mathrm{AC}_1,t}^{(60\,\mathrm{ticks})}$ are both consistent for $\overline{IV}$ under the null hypothesis, whereas only the latter is consistent under a particular class of alternatives. From Tables 3 and 4 we see that the pairs, (25,60), (30,60) and (60,180), result in very few rejections (some rejections are to be expected by pure chance). So there is little evidence that $RV_{\mathrm{AC}_1}^{(x\,\mathrm{ticks})}$ is biased for $x \geq 25$, however at the high frequencies, where $x \leq 10$, we see a large number of rejections. This shows that the implications of H_0 (Assumption 2.i) do not hold when sampling at a high frequency for both transaction and quotation.

In Figure 4, we present (signature) plots of the $t_{(m,60\,\mathrm{ticks})}$-statistics for two equities: AA and MSFT. These plots are representative for most of the 30 equities, where $t_{(m,\tilde{m})}$ is typically positive at the medium-high frequencies, whereas $t_{(m,\tilde{m})} < 0$ at the ultra-high frequencies. One possible explanation for this phenomenon is that autocorrelation in u creates the upwards bias at the medium-high frequencies, while a short-lived negative correlation between u and y^* results in a negative bias that dominates the other effect at the ultra-high frequencies. Rounding errors may explain part of these findings, but further analysis of this aspect, which is beyond the scope of this paper, is necessary to verify this explanation.

[FIGURE 4 ABOUT HERE]

4. Summary and Concluding Remarks

We have analyzed the properties of market microstructure noise and its influence on empirical measures of volatility. Our comparison of the bias corrected estimator, $RV_{\mathrm{AC}_1}^{(m)}$, of Zhou (1996) to the standard measure of realized variance revealed a substantial improvement in the precision, as the theoretical reduction of the MSE is about 50%-75%. These gains were achieved with a simple bias correction that incorporates the first-order autocovariance, and additional improvements are possible with more sophisticated corrections of the realized variance. For example, the subsample estimator of Zhang et al. (2003) is an estimator that has better asymptotic properties than $RV_{\mathrm{AC}_1}^{(m)}$ under Assumptions 1 and 2. Yet, $RV_{\mathrm{AC}_1}^{(m)}$ has some attractive properties that are useful for studying market microstructure noise. For example, if there is a short-lived time-dependence in the noise process,

then $RV^{(m)}_{\mathrm{AC}_1}$ is biased if m is above a certain threshold, but remains unbiased for small m (when the intraday returns span a period of time that is longer than the time-dependence in the noise process).

We have also evaluation of the accuracy of distributional results that are based on an assumption that there is no market microstructure noise. We showed that a 'no-noise' confidence interval based on the results of Barndorff-Nielsen & Shephard (2002) provide a reasonable accurate approximation when intraday returns are sampled at low frequencies, such as 20-minute sampling. However, when intraday returns are sampled at higher frequencies the 'no-noise' approximation are likely to be quite poor. The analogous 'no-noise' confidence interval about $RV^{(m)}_{\mathrm{AC}_1}$ yields a more accurate approximation than that of $RV^{(m)}$, but both are very misleading when intraday returns are sampled at high frequencies.

More importantly. Our empirical analysis uncovered several characteristics of market microstructure noise, where the most notably features are: (1) the noise process is time-dependent and (2) the noise process is correlated with the innovations in the efficient price process. These results were established for both transaction data and quotation data and were found to hold for intraday returns that are based on both calendar-time sampling and tick-time sampling.

Several results in the existing literature that analyze volatility estimation from high-frequency data that are contaminated with market microstructure noise, including our theoretical results in Section 2, have assumed that the noise process is independent of the efficient price and uncorrelated in time. Our empirical results suggest that the implications of these assumptions may hold (at least approximately) when intraday returns are sampled at relatively low frequencies. So the conclusions of these papers may hold as long as intraday returns are not sampled more frequently than every 15 ticks, say.

Our empirical results have shown that the sampling of intraday returns at ultra high frequencies, such as every few ticks, necessitate more general assumptions about the dependence structure of market microstructure noise. Some results are established in Hansen & Lunde (2004) who use a general specification for the noise process that can accommodate both types of dependencies that we found to be empirically relevant. We believe that interesting future research topics include: (1) deriving estimators that are robust to the various forms of dependencies; (2) a study of the robustness of existing estimators (to these forms of dependencies), such as the moving-average based estimator of Ebens (1999) and Andersen et al. (2001), and the subsample based estimator of Zhang et al. (2003). We leave this for future research.

Acknowledgements

We thank seminar participants at University of Copenhagen, University of Aarhus, Nuffield College, Carnegie Mellon University and the Econometric Forecasting and High-Frequency Data Analysis Symposium, jointly organized by Institute for Mathematical Sciences, National University of Singapore and School of Economics and Social Sciences, Singapore Management University for many valuable comments. We are particularly grateful to Per Mykland, Neil Shephard, two anonymous referees, and Torben Andersen (the editor) for their suggestions that improved this manuscript. All errors remain our responsibility. Financial support from the Danish Research Agency, grant no. 24-00-0363 is gratefully acknowledged.

Appendix of Proofs

As stated earlier, we condition on $\{\sigma^2(t)\}$ in our analysis, thus without loss of generality we treat $\sigma^2(t)$ as a deterministic function in our derivations.

Proof of Lemma 1. The bias follows directly from the decomposition $y_{i,m}^2 = y_{i,m}^{*2} + e_{i,m}^2 + 2y_{i,m}^* e_{i,m}$, since $E(e_{i,m}^2) = E(u_{i,m} - u_{i-1,m})^2 = E(u_{i,m}^2) + E(u_{i-1,m}^2) - 2E(u_{i,m}u_{i-1,m}) = 2\omega^2$, where we have used Assumption 2.*i-ii*. Similarly, we see that

$$\text{var}(RV^{(m)}) = \text{var}(\sum_{i=1}^m y_{i,m}^{*2}) + \text{var}(\sum_{i=1}^m e_{i,m}^2) + 4\,\text{var}(\sum_{i=1}^m y_{i,m}^* e_{i,m})$$

because the three sums are uncorrelated. The first sum involves uncorrelated terms such that $\text{var}(\sum_{i=1}^m y_{i,m}^{*2}) = \sum_{i=1}^m \text{var}(y_{i,m}^{*2}) = 2\sum_{i=1}^m \sigma_{i,m}^4$, where the last equality follows from the Gaussian assumption. For the second sum we find

$$\begin{aligned}
E(e_{i,m}^4) &= E(u_{i,m} - u_{i-1,m})^4 = E(u_{i,m}^2 + u_{i-1,m}^2 - 2u_{i,m}u_{i-1,m})^2 \\
&= E(u_{i,m}^4 + u_{i-1,m}^4 + 4u_{i,m}^2u_{i-1,m}^2 + 2u_{i,m}^2u_{i-1,m}^2) + 0 = 2\mu_4 + 6\omega^4, \\
E(e_{i,m}^2e_{i+1,m}^2) &= E(u_{i,m} - u_{i-1,m})^2(u_{i+1,m} - u_{i,m})^2 \\
&= E(u_{i,m}^2 + u_{i-1,m}^2 - 2u_{i,m}u_{i-1,m})(u_{i+1,m}^2 + u_{i,m}^2 - 2u_{i+1,m}u_{i,m}) \\
&= E(u_{i,m}^2 + u_{i-1,m}^2)(u_{i+1,m}^2 + u_{i,m}^2) + 0 = \mu_4 + 3\omega^4,
\end{aligned}$$

where we have used Assumption 2.*i-iii*. Thus $\text{var}(e_{i,m}^2) = 2\mu_4 + 6\omega^4 - [E(e_{i,m}^2)]^2 = 2\mu_4 + 2\omega^4$ and $\text{cov}(e_{i,m}^2, e_{i+1,m}^2) = \mu_4 - \omega^4$. Since $\text{cov}(e_{i,m}^2, e_{i+h,m}^2) = 0$ for $|h| \geq 2$ it follows that

$$\begin{aligned}
\text{var}(\textstyle\sum_{i=1}^m e_{i,m}^2) &= \textstyle\sum_{i=1}^m \text{var}(e_{i,m}^2) + \sum_{\substack{i,j=1 \\ i\neq j}}^m \text{cov}(e_{i,m}^2, e_{i+h,m}^2) \\
&= m(2\mu_4 + 2\omega^4) + 2(m-1)(\mu_4 - \omega^4) = 4m\,\mu_4 - 2(\mu_4 - \omega^4).
\end{aligned}$$

The last sum involves uncorrelated terms such that

$$\text{var}(\sum_{i=1}^m e_{i,m}y_{i,m}^*) = \sum_{i=1}^m \text{var}(e_{i,m}y_{i,m}^*) = 2\omega^2 \sum_{i=1}^m \sigma_{i,m}^2.$$

By the substitution $3\kappa = \mu_4$ we obtain the expression for the variance. The asymptotic normality is proven by Zhang et al. (2003) with an argument that is similar to that we use for $RV^{(m)}_{\mathrm{AC}_1}$ in our proof of Lemma 2, and using that $2\sum_{i=1}^{m}\sigma^4_{i,m} + 2\omega^2\sum_{i=1}^{m}\sigma^2_{i,m} - 4\omega^4 = O(1)$. ■

Proof of Lemma 2. First we note that $RV^{(m)}_{\mathrm{AC}_1} = \sum_{i=1}^{m} Y_{i,m} + U_{i,m} + V_{i,m} + W_{i,m}$, where

$$\begin{aligned} Y_{i,m} &\equiv y^*_{i,m}(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m}), \\ U_{i,m} &\equiv (u_{i,m} - u_{i-1,m})(u_{i+1,m} - u_{i-2,m}), \\ V_{i,m} &\equiv y^*_{i,m}(u_{i+1,m} - u_{i-2,m}), \\ W_{i,m} &\equiv (u_{i,m} - u_{i-1,m})(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m}), \end{aligned}$$

since $y_{i,m}(y_{i-1,m} + y_{i,m} + y_{i+1,m}) = (y^*_{i,m} + u_{i,m} - u_{i-1,m})(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m} + u_{i+1,m} - u_{i-2,m})$ $= Y_{i,m} + U_{i,m} + V_{i,m} + W_{i,m}$. Thus the properties of $RV^{(m)}_{\mathrm{AC}_1}$ are given from those of $Y_{i,m}$, $U_{i,m}$, $V_{i,m}$, and $W_{i,m}$. Given Assumptions 1 and 2.i, it follows directly that $E(Y_{i,m}) = \sigma^2_{i,m}$, and $E(U_{i,m}) = E(V_{i,m}) = E(W_{i,m}) = 0$, which shows that $E[RV^{(m)}_{\mathrm{AC}_1}] = \sum_{i=1}^{m}\sigma^2_{i,m}$. Note that $E(U_{i,m})$ consists of terms $E(u_{i,m}u_{j,m})$ where $i \neq j$ so Assumption 2.i suffices to establish that the expected value is zero. Given Assumptions 1 and 2.i-ii, the variance of $RV^{(m)}_{\mathrm{AC}_1}$ is given by

$$\mathrm{var}[RV^{(m)}_{\mathrm{AC}_1}] = \mathrm{var}[\sum_{i=1}^{m} Y_{i,m} + U_{i,m} + V_{i,m} + W_{i,m}] = (1) + (2) + (3) + (4) + (5),$$

where $(1) = \mathrm{var}(\sum_{i=1}^{m} Y_{i,m})$, $(2) = \mathrm{var}(\sum_{i=1}^{m} U_{i,m})$, $(3) = \mathrm{var}(\sum_{i=1}^{m} V_{i,m})$, $(4) = \mathrm{var}(\sum_{i=1}^{m} W_{i,m})$, $(5) = 2\mathrm{cov}(\sum_{i=1}^{m} V_{i,m}, \sum_{i=1}^{m} W_{i,m})$, since all other sums are uncorrelated. Next, we derive the expressions of each of these five terms.

1. $Y_{i,m} = y^*_{i,m}(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m})$ and given Assumption 1 it follows that $E[y^{*2}_{i,m}y^{*2}_{j,m}] = \sigma^2_{i,m}\sigma^2_{j,m}$ for $i \neq j$, and $E[y^{*2}_{i,m}y^{*2}_{j,m}] = E[y^{*4}_{i,m}] = 3\sigma^4_{i,m}$ for $i = j$, such that

$$\mathrm{var}(Y_{i,m}) = 3\sigma^4_{i,m} + \sigma^2_{i,m}\sigma^2_{i-1,m} + \sigma^2_{i,m}\sigma^2_{i+1,m} - [\sigma^2_{i,m}]^2 = 2\sigma^4_{i,m} + \sigma^2_{i,m}\sigma^2_{i-1,m} + \sigma^2_{i,m}\sigma^2_{i+1,m}.$$

The first-order autocorrelation of $Y_{i,m}$ is

$$\begin{aligned} E[Y_{i,m}Y_{i+1,m}] &= E[y^*_{i,m}(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m})y^*_{i+1,m}(y^*_{i,m} + y^*_{i+1,m} + y^*_{i+2,m})] \\ &= E[y^*_{i,m}(y^*_{i,m} + y^*_{i+1,m})y^*_{i+1,m}(y^*_{i,m} + y^*_{i+1,m})] + 0 \\ &= 2E[y^{*2}_{i,m}y^{*2}_{i+1,m}] = 2\sigma^2_{i,m}\sigma^2_{i+1,m}, \end{aligned}$$

such that $\mathrm{cov}(Y_{i,m}, Y_{i+1,m}) = \sigma^2_{i,m}\sigma^2_{i+1,m}$, whereas $\mathrm{cov}(Y_{i,m}, Y_{i+h,m}) = 0$ for $|h| \geq 2$. Thus

$$\begin{aligned} (1) &= \sum_{i=1}^{m}(2\sigma^4_{i,m} + \sigma^2_{i,m}\sigma^2_{i-1,m} + \sigma^2_{i,m}\sigma^2_{i+1,m}) + \sum_{i=2}^{m}\sigma^2_{i,m}\sigma^2_{i-1,m} + \sum_{i=1}^{m-1}\sigma^2_{i,m}\sigma^2_{i+1,m} \\ &= 2\sum_{i=1}^{m}\sigma^4_{i,m} + 2\sum_{i=1}^{m}\sigma^2_{i,m}\sigma^2_{i-1,m} + 2\sum_{i=1}^{m}\sigma^2_{i,m}\sigma^2_{i+1,m} - \sigma^2_{1,m}\sigma^2_{0,m} - \sigma^2_{m,m}\sigma^2_{m+1,m} \\ &= 6\sum_{i=1}^{m}\sigma^4_{i,m} - 2\sum_{i=1}^{m}\sigma^2_{i,m}(\sigma^2_{i,m} - \sigma^2_{i-1,m}) + 2\sum_{i=1}^{m}\sigma^2_{i,m}(\sigma^2_{i+1,m} - \sigma^2_{i,m}) \\ &\quad -\sigma^2_{1,m}\sigma^2_{0,m} - \sigma^2_{m,m}\sigma^2_{m+1,m} \end{aligned}$$

$$
\begin{aligned}
&= 6\sum_{i=1}^{m}\sigma_{i,m}^4 - 2\sum_{i=2}^{m}\sigma_{i,m}^2(\sigma_{i,m}^2-\sigma_{i-1,m}^2) + 2\sum_{i=1}^{m-1}\sigma_{i,m}^2(\sigma_{i+1,m}^2-\sigma_{i,m}^2)\\
&\quad -\sigma_{1,m}^2\sigma_{0,m}^2 - \sigma_{m,m}^2\sigma_{m+1,m}^2 - 2\sigma_{1,m}^2(\sigma_{1,m}^2-\sigma_{0,m}^2) + 2\sigma_{m,m}^2(\sigma_{m+1,m}^2-\sigma_{m,m}^2)\\
&= 6\sum_{i=1}^{m}\sigma_{i,m}^4 - 2\sum_{i=1}^{m-1}(\sigma_{i+1,m}^2-\sigma_{i,m}^2)^2 - 2(\sigma_{1,m}^4+\sigma_{m,m}^4) + \sigma_{1,m}^2\sigma_{0,m}^2 + \sigma_{m,m}^2\sigma_{m+1,m}^2
\end{aligned}
$$

2. $U_{i,m} = (u_{i,m}-u_{i-1,m})(u_{i+1,m}-u_{i-2,m})$ and from $E(U_{i,m}^2) = E(u_{i,m}-u_{i-1,m})^2E(u_{i+1,m}-u_{i-2,m})^2$ it follows that $\text{var}(U_{i,m}) = 4\omega^4$. The first and second order autocovariance are given by

$$
\begin{aligned}
E(U_{i,m}U_{i+1,m}) &= E[(u_{i,m}-u_{i-1,m})(u_{i+1,m}-u_{i-2,m})(u_{i+1,m}-u_{i,m})(u_{i+2,m}-u_{i-1,m})]\\
&= E[u_{i-1,m}u_{i+1,m}u_{i+1,m}u_{i-1,m}] + 0 = \omega^4, \qquad \text{and}\\
E(U_{i,m}U_{i+2,m}) &= E[(u_{i,m}-u_{i-1,m})(u_{i+1,m}-u_{i-2,m})(u_{i+2,m}-u_{i+1,m})(u_{i+3,m}-u_{i,m})]\\
&= E[u_{i,m}u_{i+1,m}u_{i+1,m}u_{i,m}] + 0 = \omega^4,
\end{aligned}
$$

whereas $E(U_{i,m}U_{i+h,m}) = 0$ for $|h| \geq 3$. Thus, $(2) = m4\omega^4 + 2(m-1)\omega^4 + 2(m-2)\omega^4 = 8\omega^4 m - 6\omega^4$.

3. $V_{i,m} = y_{i,m}^*(u_{i+1,m}-u_{i-2,m})$ such that $\text{var}(V_{i,m}^2) = \sigma_{i,m}^2 2\omega^2$ and $E[V_{i,m}V_{i+h,m}] = 0$ for all $h \neq 0$. Thus $(3) = \text{var}(\sum_{i=1}^m V_{i,m}) = 2\omega^2\sum_{i=1}^m\sigma_{i,m}^2$.

4. $W_{i,m} = (u_{i,m}-u_{i-1,m})(y_{i-1,m}^* + y_{i,m}^* + y_{i+1,m}^*)$ such that $\text{var}(W_{i,m}^2) = 2\omega^2(\sigma_{i-1,m}^2+\sigma_{i,m}^2+\sigma_{i+1,m}^2)$. The first order autocovariance equals

$$
\text{cov}(W_{i,m}, W_{i+1,m}) = E[-u_{i,m}^2(y_{i,m}^{*2} + y_{i+1,m}^{*2})] = -\omega^2(\sigma_{i,m}^2+\sigma_{i+1,m}^2),
$$

while $\text{cov}(W_{i,m}, W_{i+h,m}) = 0$ for $|h| \geq 2$. Thus

$$
\begin{aligned}
(4) &= \sum_{i=1}^{m}[2\omega^2(\sigma_{i-1,m}^2+\sigma_{i,m}^2+\sigma_{i+1,m}^2) - \sum_{i=2}^{m}\omega^2(\sigma_{i,m}^2+\sigma_{i-1,m}^2) - \sum_{i=1}^{m-1}\omega^2(\sigma_{i,m}^2+\sigma_{i+1,m}^2)]\\
&= \omega^2\sum_{i=1}^{m}(\sigma_{i-1,m}^2+\sigma_{i+1,m}^2) + \omega^2[\sigma_{1,m}^2+\sigma_{0,m}^2+\sigma_{m,m}^2+\sigma_{m+1,m}^2]\\
&= 2\omega^2\sum_{i=1}^{m}\sigma_{i,m}^2 + \omega^2[\sigma_{0,m}^2-\sigma_{m,m}^2+\sigma_{m+1,m}^2-\sigma_{1,m}^2] + \omega^2[\sigma_{1,m}^2+\sigma_{0,m}^2+\sigma_{m,m}^2+\sigma_{m+1,m}^2]\\
&= 2\omega^2\sum_{i=1}^{m}\sigma_{i,m}^2 + 2\omega^2[\sigma_{0,m}^2+\sigma_{m+1,m}^2].
\end{aligned}
$$

5. The autocovariances between the last two terms are given by

$$
E[V_{i,m}W_{i+h,m}] = E[y_{i,m}^*(u_{i+1,m}-u_{i-2,m})(u_{i+h,m}-u_{i-1+h,m})(y_{i-1+h,m}^* + y_{i+h,m}^* + y_{i+1+h,m}^*)],
$$

showing that $\text{cov}(V_{i,m}, W_{i\pm1,m}) = \omega^2\sigma_{i,m}^2$, while all other covariances are zero. From this we conclude that $(5) = 2[2\sum_{i=1}^m\omega^2\sigma_{i,m}^2 - \omega^2(\sigma_{1,m}^2+\sigma_{m,m}^2)] = 4\omega^2\sum_{i=1}^m\sigma_{i,m}^2 - 2\omega^2(\sigma_{1,m}^2+\sigma_{m,m}^2)$.

By adding up the five terms we find

$$
6\sum_{i=1}^{m}\sigma_{i,m}^4 - 2\sum_{i=1}^{m-1}(\sigma_{i+1,m}^2-\sigma_{i,m}^2)^2 - 2(\sigma_{1,m}^4+\sigma_{m,m}^4) + \sigma_{1,m}^2\sigma_{0,m}^2 + \sigma_{m,m}^2\sigma_{m+1,m}^2 + 8\omega^4 m - 6\omega^4
$$

$$+ 2\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2 + 2\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2 + 2\omega^2[\sigma_{0,m}^2 + \sigma_{m+1,m}^2] + 4\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2 - 2\omega^2[\sigma_{1,m}^2 + \sigma_{m,m}^2]$$
$$= 8\omega^4 m + 8\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2 - 6\omega^4 + 6\sum_{i=1}^{m} \sigma_{i,m}^4 + r_m,$$

where

$$\begin{aligned} r_m \quad \equiv \quad & -2\sum_{i=1}^{m}(\sigma_{i+1,m}^2 - \sigma_{i,m}^2)^2 - 2(\sigma_{1,m}^4 + \sigma_{m,m}^4) + \sigma_{1,m}^2\sigma_{0,m}^2 + \sigma_{m,m}^2\sigma_{m+1,m}^2 \\ & +2\omega^2(\sigma_{0,m}^2 - \sigma_{1,m}^2 + \sigma_{m+1,m}^2 - \sigma_{m,m}^2). \end{aligned}$$

Under BTS it follows immediately that $r_m = O(m^{-2})$. Under CTS we use the Lipschitz condition, which states that $\exists \epsilon > 0$ such that $|\sigma^2(t) - \sigma^2(t+h)| \leq \epsilon h$ for all t and all h. This shows that $|\sigma_{i,m}^2| = |\int_{t_{i-1,m}}^{t_{i,m}} \sigma^2(s)ds| \leq \delta \sup_{t_{i-1,m} \leq s \leq t_{i,m}} \sigma^2(s) = O(m^{-1})$, since $\delta = \delta_{i,m} = (b-a)/m = O(m^{-1})$ under CTS, and

$$\begin{aligned} |\sigma_{i,m}^2 - \sigma_{i-1,m}^2| \quad &= \quad |\int_{t_{i-1,m}}^{t_{i,m}} \sigma^2(s) - \sigma^2(s-\delta)ds| \leq \int_{t_{i-1,m}}^{t_{i,m}} |\sigma^2(s) - \sigma^2(s-\delta)|ds \\ &\leq \quad \delta \sup_{t_{i-1,m} \leq s \leq t_{i,m}} |\sigma^2(s) - \sigma^2(s-\delta)| \leq \delta^2\epsilon = O(m^{-2}). \end{aligned}$$

Thus $\sum_{i=1}^{m}(\sigma_{i+1,m}^2 - \sigma_{i,m}^2)^2 \leq m \cdot (\delta\frac{\epsilon}{m})^2 = O(m^{-3})$, which proves that $r_m = O(m^{-2})$, under CTS.

The asymptotic normality is established by expressing $RV_{\mathrm{AC}_1}^{(m)}$ as a sum of a martingale difference sequence. Let $u_{i,m} = u(t_{i,m})$ and define the sigma algebra $\mathcal{F}_{i,m} = \sigma(y_{i,m}^*, y_{i-1,m}^*, \ldots, u_{i,m}, u_{i-1,m}, \ldots)$. First note that $y_{i,m}(y_{i-1,m}+y_{i,m}+y_{i+1,m}) = \sigma_{i,m}^2 + \xi_{i-1,m}^{(1)} + \xi_{i,m}^{(2)} + \xi_{i+1,m}^{(3)}$, where

$$\begin{aligned} \xi_{i-1,m}^{(1)} \quad &\equiv \quad -u_{i-1,m}y_{i-1,m}^* + u_{i-1,m}u_{i-2,m}, \\ \xi_{i,m}^{(2)} \quad &\equiv \quad y_{i,m}^*y_{i,m}^* - \sigma_{i,m}^2 + y_{i,m}^*y_{i-1,m}^* - y_{i,m}^*u_{i-2,m} + u_{i,m}y_{i-1,m}^* + u_{i,m}y_{i,m}^* - u_{i,m}u_{i-2,m} - u_{i-1,m}y_{i,m}^*, \\ \xi_{i+1,m}^{(3)} \quad &\equiv \quad y_{i,m}^*y_{i+1,m}^* + y_{i,m}^*u_{i+1,m} + u_{i,m}y_{i+1,m}^* + u_{i,m}u_{i+1,m} - u_{i-1,m}y_{i+1,m}^* - u_{i-1,m}u_{i+1,m}. \end{aligned}$$

So if we define $\xi_{i,m} \equiv (\xi_{i,m}^{(1)} + \xi_{i,m}^{(2)} + \xi_{i,m}^{(3)})$ (using the conventions $\xi_{0,m}^{(2)} = \xi_{0,m}^{(3)} = \xi_{1,m}^{(3)} = \xi_{m,m}^{(1)} = \xi_{m+1,m}^{(1)} = \xi_{m+1,m}^{(2)} = 0$), it follows that $[RV_{\mathrm{AC}_1}^{(m)} - IV] = \sum_{i=0}^{m+1} \xi_{i,m}$, where $\{\xi_{i,m}, \mathcal{F}_{i,m}\}_{i=0}^{m+1}$ is a martingale difference sequence that is squared integrable, since

$$E(\xi_{i,m}^2) = \begin{cases} \omega^2\sigma_0^2 + \omega^4 < \infty, & \text{for } i = 0, \\ 2\omega^4 + \sigma_0^2\sigma_1^2 + \omega^2\sigma_0^2 + 2\omega^2\sigma_1^2 + \sigma_1^4 < \infty, & \text{for } i = 1, \\ 2\sigma_{i,m}^4 + 4\sigma_{i,m}^2\sigma_{i-1,m}^2 + 4\sigma_{i,m}^2\omega^2 + 4\sigma_{i-1,m}^2\omega^2 + 8\omega^4 < \infty, & \text{for } 1 < i < m, \\ \sigma_m^4 + 4\sigma_m^2\sigma_{m-1}^2 + 6\omega^2\sigma_m^2 + 4\omega^2\sigma_{m-1}^2 + 5\omega^4 < \infty, & \text{for } i = m, \\ \sigma_m^2\sigma_{m+1}^2 + 2\omega^2\sigma_{m+1}^2 + 2\omega^4 < \infty, & \text{for } i = m+1, \end{cases}$$

Since $m^{-1/2}[RV_{\mathrm{AC}_1}^{(m)} - IV] = m^{-1/2}\sum_{i=0}^{m+1} \xi_{i,m}$, we can apply the central limit theorem for squared integrable martingales, see Shiryaev (1995, p. 543, theorem 4), where the only remaining condition to be verified, is the conditional Lindeberg condition:

$$\sum_{i=0}^{m+1} E\left[m^{-1}\xi_{i,m}^2 1_{\{|m^{-1/2}\xi_{i,m}| > \varepsilon\}} | \mathcal{F}_{i-1,m}\right] \xrightarrow{p} 0, \qquad \text{as } m \to \infty.$$

Since $E[\xi_{i,m}^2 1_{\{|m^{-1/2}\xi_{i,m}|>\varepsilon\}}]$ is bounded by $E[\xi_{i,m}^2] \leq \infty$ and $\sup_i P(1_{\{|\xi_{i,m}|>\varepsilon\sqrt{m}\}} = 0) \to 0$, for all $\varepsilon > 0$, it follows that

$$E\left|m^{-1}\sum_{i=0}^{m+1}\xi_{i,m}^2 1_{\{|m^{-1/2}\xi_{i,m}|>\varepsilon\}} - 0\right| \leq m^{-1}\sum_{i=0}^{m+1}\left|E[\xi_{i,m}^2 1_{\{|m^{-1/2}\xi_{i,m}|>\varepsilon\}}] - 0\right| \to 0, \quad \text{as } m \to \infty.$$

The Lindeberg condition now follows because convergence in $\mathcal{L}_1$ implies convergence in probability. ■

Proof of Corollary 3. The MSE's are given from Lemmas 1 and 2, since BTS implies that

$$r_m = 0 - 2(IV^2/m^2 + IV^2/m^2) + IV^2/m^2 + IV^2/m^2 + 0 = -2IV^2/m^2.$$

Equating $\partial \text{MSE}(RV^{(m)})/\partial m \propto 4\lambda^2 m + 6\lambda^2 - m^{-2}$ with zero yields the first order condition of the corollary, and the second result follows similarly from $\partial \text{MSE}(RV_{\text{AC}_1}^{(m)})/\partial m \propto 4\lambda^2 - 3m^{-2} + 2m^{-3}$. ■

References

Andersen, T. G. & Bollerslev, T. (1997), 'Intraday periodicity and volatility persistence in financial markets', *Journal of Empirical Finance* **4**, 115–158.

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

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2000), 'Great realizations', *Risk* **13**(3), 105–108.

Andreou, E. & Ghysels, E. (2002), 'Rolling-sample volatility estimators: Some new theoretical, simulation, and empirical results', *Journal of Business & Economic Statistics* **20**(3), 363–376.

Awartani, B., Corradi, V. & Distaso, W. (2004), Testing and modelling market microstructure effects with an application to the dow jones industrial average. Unpublished Manuscript, Queen Mary, University of London and University of Exeter.

Bandi, F. M. & Russell, J. R. (2003), Microstructure noise, realized volatility, and optimal sampling, Working paper, Graduate School of Business, The University of Chicago.

Barndorff-Nielsen, O. E. & Shephard, N. (2002), 'Econometric analysis of realised volatility and its use in estimating stochastic volatility models', *Journal of the Royal Statistical Society* **B 64**, 253–280.

Barndorff-Nielsen, O. E. & Shephard, N. (2003), 'Realized power variation and stochastic volatility', *Bernoulli* **9**, 243–265.

Billingsley, P. (1995), *Probability and Measure*, 3nd edn, John Wiley and Sons, New York.

Bollen, B. & Inder, B. (2002), 'Estimating daily volatility in financial markets utilizing intraday data', *Journal of Empirical Finance* **9**, 551–562.

Corsi, F., Zumbach, G., Müller, U. & Dacorogna, M. (2001), 'Consistent high-precision volatility from high-frequency data', *Economic Notes* **30**(2), 183–204.

Curci, G. & Corsi, F. (2004), Discrete sine transform approach for realized volatility measurement. Unpublished Manuscript, Research Paper, University of Southern Switzerland.

Dacorogna, M. M., Gencay, R., Müller, U., Olsen, R. B. & Pictet, O. V. (2001), *An Introduction to High-Frequency Finance*, Academic Press, London.

Ebens, H. (1999), Realized stock volatility, Working paper 420, Johns Hopkins University.

French, K. R., Schwert, G. W. & Stambaugh, R. F. (1987), 'Expected stock returns and volatility', *Journal of Financial Economics* **19**(1), 3–29.

Hansen, P. R. & Lunde, A. (2003), 'An optimal and unbiased measure of realized variance based on intermittent high-frequency data'. Mimeo prepared for the CIREQ-CIRANO Conference: Realized Volatility. Montreal, November 2003.

Hansen, P. R. & Lunde, A. (2004), 'An unbiased measure of realized variance'. Brown University Working Paper, 2004
`http://www.econ.brown.edu/fac/Peter_Hansen`.

Meddahi, N. (2002), 'A theoretical comparison between integrated and realized volatility', *Journal of Applied Econometrics* **17**, 479–508.

Müller, U. A., Dacorogna, M. M., Olsen, R. B., Pictet, O. V., Schwarz, M. & Morgenegg, C. (1990), 'Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis', *Journal of Banking and Finance* **14**, 1189–1208.

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

Oomen, R. A. C. (2002), 'Modelling realized variance when returns are serially correlated'. manuscript, Warwick Business School, The University of Warwick.

Oomen, R. A. C. (2004*a*), 'Properties of bias corrected realized variance in calendar time and business time'. manuscript, Warwick Business School, The University of Warwick.

Oomen, R. A. C. (2004*b*), 'Properties of realized variance for a pure jump process: Calendar time sampling versus business time sampling'. manuscript, Warwick Business School, The University of Warwick.

Shiryaev, A. N. (1995), *Probability*, 2nd edn, Springer-Verlag, New York.

Wasserfallen, W. & Zimmermann, H. (1985), 'The behavior of intraday exchange rates', *Journal of Banking and Finance* **9**, 55–72.

Zhang, L., Mykland, P. A. & Aït-Sahalia, Y. (2002), A tale of two time scales: Determining integrated volatility with noisy high frequency data?, Tech. rep. university of chicago.

Zhang, L., Mykland, P. A. & Aït-Sahalia, Y. (2003), A tale of two time scales: Determining integrated volatility with noisy high frequency data, Working Paper w10111, NBER.

Zhou, B. (1996), 'High-frequency data and volatility in foreign-exchange rates', *Journal of Business & Economic Statistics* **14**(1), 45–52.

Table 1: Equities included in our empirical analysis.

			Trans./day		Quotes/day		
Symbol	Name	Exchange	All	$\#\Delta p \neq 0$	All	$\#\Delta p \neq 0$	#Days
AA	ALCOA INC	NYSE	1454	699	2007	770	742
AXP	AMERICAN EXPRESS	NYSE	2267	1179	2938	1267	741
BA	BOEING COMPANY	NYSE	1687	817	2529	1040	741
C	CITIGROUP	NYSE	3236	1578	4032	1495	742
CAT	CATERPILLAR INC	NYSE	1218	601	1955	781	742
DD	DU PONT DE NEMOURS	NYSE	1727	905	2716	1030	741
DIS	WALT DISNEY	NYSE	2049	1047	3026	1087	742
EK	EASTMAN KODAK	NYSE	1089	501	1800	702	741
GE	GENERAL ELECTRIC	NYSE	3690	1859	4014	1655	742
GM	GENERAL MOTORS	NYSE	1591	755	2603	991	742
HD	HOME DEPOT INC	NYSE	2710	1337	3086	1289	742
HON	HONEYWELL	NYSE	1477	677	2292	930	741
HPQ	HEWLETT-PACKARD	NYSE	2246	1053	2918	1194	742
IBM	INT. BUSINESS MACHINES	NYSE	3138	1901	4883	2249	741
INTC	INTEL CORP	NASDAQ	15907	12388	13579	4566	750
IP	INTERNATIONAL PAPER	NYSE	1489	710	2269	836	742
JNJ	JOHNSON AND JOHNSON	NYSE	2023	1139	2377	998	742
JPM	J.P. MORGAN	NYSE	2280	1188	3091	1364	742
KO	COCA-COLA	NYSE	1788	904	2298	948	742
MCD	MCDONALDS	NYSE	1690	797	2313	807	742
MMM	MINNESOTA MNG MFG	NYSE	1487	803	2267	1094	741
MO	PHILIP MORRIS	NYSE	2080	1055	3254	1009	741
MRK	MERCK	NYSE	2117	1032	2549	1064	742
MSFT	MICROSOFT	NASDAQ	15324	11947	13257	6246	747
PG	PROCTER & GAMBLE	NYSE	2004	1017	3409	1299	742
SBC	SBC COMMUNICATIONS	NYSE	2293	1108	2902	1106	741
T	AT & T CORP	NYSE	2104	789	2267	734	742
UTX	UNITED TECHNOLOGIES	NYSE	1360	687	2112	953	741
WMT	WAL-MART STORES	NYSE	2380	1261	3084	1226	742
XOM	EXXON MOBIL	NYSE	2463	1253	3104	1176	741

The table lists the equities used in our empirical analysis. For each equity, we extract data from the exchange where it is most actively traded (third column). The average number of transactions and quotes per day are given in columns 4-7. The final columns report the number of trading day for each asset.

Table 2: Noise-to-signal ratio, optimal sampling.

	Trades						Quotes	
Asset	$\hat{\omega}^2 \cdot 100$	$\overline{IV}$	$\hat{\lambda} \cdot 100$	m_0^*	m_1^*	ΔMSE	$\hat{\omega}^2 \cdot 100$	$\overline{IV}$
AA	0.4217	5.797	0.0727	77	1190	64.5%	-0.0966	5.855
AXP	0.1182	5.944	0.0199	184	4354	75.7%	-0.0870	5.953
BA	0.2356	5.238	0.0450	107	1924	69.1%	-0.0690	5.011
C	0.1814	5.138	0.0353	126	2453	71.2%	-0.0471	5.172
CAT	0.2238	4.909	0.0456	106	1899	69.0%	-0.1156	4.818
DD	0.2697	4.716	0.0572	91	1514	66.9%	-0.0778	4.735
DIS	0.6089	5.770	0.1055	60	820	60.7%	-0.0511	5.614
EK	0.0838	5.125	0.0164	210	5294	77.1%	-0.1047	5.121
GE	0.1964	4.619	0.0425	111	2036	69.6%	-0.0347	4.484
GM	-0.0169	4.700	-0.0036	n/a	n/a	n/a	-0.1346	4.694
HD	0.2728	5.788	0.0471	104	1837	68.7%	-0.0906	5.689
HON	0.2304	6.926	0.0333	131	2603	71.7%	-0.1215	7.037
HPQ	0.2949	10.08	0.0292	142	2961	72.8%	-0.1132	9.455
IBM	0.0679	4.515	0.0150	222	5759	77.7%	-0.0305	4.456
INTC	0.2455	10.91	0.0225	170	3847	74.8%	-0.1843	10.77
IP	0.4114	5.463	0.0753	76	1149	64.2%	-0.1256	5.559
JNJ	0.1088	2.465	0.0441	108	1961	69.3%	-0.0393	2.341
JPM	0.1008	7.389	0.0136	237	6348	78.3%	-0.1018	7.077
KO	0.1847	3.203	0.0577	90	1501	66.8%	-0.0656	3.280
MCD	0.6553	3.862	0.1697	44	510	55.3%	-0.0168	3.952
MMM	-0.0087	3.389	-0.0026	n/a	n/a	n/a	-0.0814	3.320
MO	0.8021	3.962	0.2025	39	427	53.1%	-0.0648	4.169
MRK	0.0964	3.446	0.0280	147	3094	73.1%	-0.0580	3.550
MSFT	0.1688	6.120	0.0276	148	3139	73.2%	-0.0888	5.872
PG	0.1423	3.063	0.0465	105	1863	68.8%	-0.0202	2.995
SBC	0.2865	4.881	0.0587	89	1475	66.6%	-0.0734	5.007
T	0.7740	5.531	0.1399	50	618	57.5%	-0.1365	5.842
UTX	-0.0038	4.875	-0.0008	n/a	n/a	n/a	-0.1070	4.977
WMT	0.1755	4.946	0.0355	125	2440	71.2%	-0.0995	5.024
XOM	0.0979	2.588	0.0378	120	2289	70.6%	-0.0341	2.532

The table presents empirical estimates of ω^2, the average IV, the noise-to-signal ratio, the optimal sampling frequencies for $RV^{(m)}$ and $RV^{(m)}_{\mathrm{AC}_1}$, and the reduction of the MSE, $100[\gamma_0^2(\hat{\lambda}, m_0^*) - \gamma_1^2(\hat{\lambda}, m_1^*)]/\gamma_0^2(\hat{\lambda}, m_0^*)$ based on transaction data. For quotation data we observe negative estimates of ω^2 which is evidence of a negative correlation between the noise process and true returns.

Table 3: Test for time-independent noise (transaction data).

Asset	$t_{(1,60)}$	$t_{(2,60)}$	$t_{(5,60)}$	$t_{(10,60)}$	$t_{(15,60)}$	$t_{(20,60)}$	$t_{(25,60)}$	$t_{(30,60)}$	$t_{(60,180)}$
AA	**-10.3**	1.07	**2.54**	**2.31**	1.00	0.11	-0.67	-0.49	-1.10
AXP	**-7.49**	**3.57**	**5.94**	**3.60**	1.88	1.04	0.84	0.26	-0.78
BA	**-7.12**	1.86	**2.41**	0.30	-1.01	-1.00	-1.06	-1.21	-0.64
C	**-4.79**	**8.65**	**7.88**	**5.19**	**3.32**	**2.69**	1.31	0.74	0.43
CAT	**-9.84**	-1.32	0.62	0.00	-1.04	-0.66	-0.94	-0.40	**2.29**
DD	**-11.1**	1.19	**2.84**	**2.20**	1.04	-0.17	-0.28	-1.08	0.32
DIS	**-15.1**	**5.87**	**4.16**	**3.73**	**2.09**	1.03	0.53	0.18	1.70
EK	**-5.59**	-0.72	1.28	1.59	1.42	0.91	0.64	1.48	1.45
GE	**-13.2**	**4.29**	**5.25**	**5.19**	**3.19**	**2.57**	1.05	1.07	-0.12
GM	**-11.0**	**-4.18**	-0.79	-1.05	-1.49	**-2.07**	-1.62	**-2.06**	-0.27
HD	**-14.0**	**3.06**	**4.22**	**3.55**	**2.15**	1.29	-0.21	0.36	-1.22
HON	**-5.30**	1.33	1.92	1.51	0.52	0.90	0.91	1.15	-0.03
HPQ	**-5.02**	**2.23**	**2.64**	0.95	0.16	0.03	-1.18	-1.43	0.53
IBM	**-3.97**	1.81	**4.01**	1.90	0.20	-0.31	-0.63	-0.82	-1.07
INTC	**-21.0**	**-13.2**	**-14.9**	**-9.78**	**-6.15**	**-4.79**	**-2.72**	-1.36	-1.03
IP	**-9.58**	1.34	**2.28**	1.87	1.31	0.88	0.85	0.78	0.71
JNJ	**-7.87**	**6.74**	**7.12**	**5.38**	**3.59**	**2.91**	**2.78**	1.95	-0.56
JPM	**-2.42**	0.60	0.26	-0.46	-1.42	-1.71	**-2.05**	**-2.24**	0.45
KO	**-10.3**	**3.99**	**4.43**	**2.79**	1.52	1.30	1.33	1.21	-0.73
MCD	**-12.9**	**5.75**	**4.35**	**3.51**	**2.13**	0.64	0.67	1.21	-1.10
MMM	-1.79	**2.10**	**2.77**	1.41	0.43	-0.07	-1.08	-0.86	1.88
MO	**-15.6**	**2.16**	-0.25	0.95	-0.48	-0.21	-0.31	-0.54	-1.92
MRK	**-9.17**	**3.47**	**5.17**	**4.35**	**2.61**	1.94	1.51	1.43	0.37
MSFT	**-19.4**	**-8.13**	**-12.8**	**-10.1**	**-7.11**	**-5.39**	**-3.65**	**-3.25**	-1.55
PG	**-5.82**	**3.20**	**3.58**	1.62	0.26	-0.44	-1.57	-0.71	-0.32
SBC	**-9.88**	**7.71**	**9.08**	**7.55**	**5.26**	**3.34**	**2.19**	1.53	-1.72
T	**-13.6**	**7.84**	**2.13**	**4.07**	1.68	0.70	-0.09	0.00	**-3.12**
UTX	**-2.19**	1.35	**3.43**	**2.73**	1.65	1.03	0.17	0.21	-0.28
WMT	**-12.5**	**3.71**	**4.14**	**2.79**	1.42	-0.08	-0.80	-1.33	-1.06
XOM	**-3.16**	**9.65**	**10.4**	**6.62**	**3.88**	**2.04**	0.37	-0.06	-1.48

This table reports t-statistics for the hypothesis that the pricing errors are independent. Bold font is identifies the statistics that are significant at the 5%-significance level. The statistic $t_{(x,60)}$, which compares $RV_{\mathrm{AC}_1}^{(x \text{ ticks})}$ to $RV_{\mathrm{AC}_1}^{(60 \text{ ticks})}$, has power against alternatives for which the dependence endures for more that x ticks.

Table 4: Test for time-independent noise (quotation data).

Asset	$t_{(1,60)}$	$t_{(2,60)}$	$t_{(5,60)}$	$t_{(10,60)}$	$t_{(15,60)}$	$t_{(20,60)}$	$t_{(25,60)}$	$t_{(30,60)}$	$t_{(60,180)}$
AA	**-5.02**	**-1.99**	1.29	1.32	0.08	-0.49	-1.15	-1.29	-0.50
AXP	**-7.34**	-1.28	**5.08**	**3.75**	**1.96**	1.16	0.07	0.01	0.08
BA	**-4.08**	0.09	**3.70**	**2.55**	0.99	0.39	-0.07	-0.32	-1.49
C	-0.01	**4.81**	**7.48**	**5.60**	**3.72**	**2.30**	1.48	0.88	-0.04
CAT	**-7.32**	**-3.70**	0.25	0.28	-0.33	-1.01	-0.57	-0.41	1.21
DD	**-5.67**	**-2.12**	**2.19**	1.87	0.68	0.04	-1.12	-1.27	0.07
DIS	-1.65	**2.46**	**5.83**	**4.51**	**2.99**	**2.06**	1.38	1.65	-0.14
EK	**-5.28**	**-2.97**	0.20	1.08	0.71	0.34	0.60	0.48	0.86
GE	**-3.59**	1.17	**6.39**	**5.76**	**3.88**	**2.57**	**2.49**	**2.11**	-0.19
GM	**-12.5**	**-8.92**	**-2.91**	-1.28	-1.39	**-2.12**	-1.62	-1.71	-1.80
HD	**-6.36**	-1.07	**4.56**	**3.73**	**2.48**	1.14	-0.21	0.25	**-2.06**
HON	**-3.95**	-1.30	1.52	1.28	0.95	0.38	0.01	0.42	-0.09
HPQ	**-3.73**	-0.22	**4.01**	**3.28**	1.61	0.65	1.19	0.45	-0.45
IBM	**-7.53**	**-2.49**	**4.40**	**3.91**	1.79	-0.12	-0.30	-0.87	-1.90
INTC	**-19.5**	**-13.3**	**-3.34**	1.25	**2.05**	1.90	1.81	1.62	**2.46**
IP	**-5.96**	**-2.88**	0.70	1.14	0.55	-0.61	-0.36	-0.57	0.24
JNJ	-1.63	**2.95**	**5.87**	**4.31**	**3.00**	**2.47**	1.86	1.69	-0.76
JPM	**-3.69**	-0.63	1.79	1.40	0.67	0.35	-0.67	-1.04	-1.44
KO	**-5.37**	-1.17	**2.83**	1.79	1.17	0.44	0.43	-0.06	0.08
MCD	-1.33	1.13	**3.42**	1.96	0.72	0.10	-0.21	-0.80	0.03
MMM	**-5.13**	-0.49	**3.36**	**2.47**	1.30	0.83	-0.10	-0.23	-0.38
MO	**-6.43**	**-3.42**	-0.81	-1.13	-1.79	**-1.98**	**-1.98**	**-2.08**	-0.29
MRK	**-6.50**	-1.83	**3.02**	**3.06**	**2.04**	0.72	0.11	-0.38	0.83
MSFT	**-18.4**	**-12.0**	-1.21	**3.54**	**4.69**	**4.85**	**4.76**	**4.31**	1.26
PG	**-2.62**	0.29	**3.81**	**3.04**	1.96	0.48	0.18	-0.50	-0.57
SBC	**-4.01**	1.83	**7.84**	**7.52**	**4.90**	**3.21**	1.38	1.01	-1.18
T	**-4.27**	0.74	**3.92**	1.64	-0.25	-0.79	-1.27	-1.10	**-2.77**
UTX	**-5.37**	-1.90	**2.11**	**3.08**	**2.49**	1.89	1.15	0.05	-0.37
WMT	**-6.38**	-1.05	**3.37**	**2.09**	-0.05	-1.18	**-2.24**	**-2.92**	-0.75
XOM	1.09	**5.98**	**8.84**	**5.81**	**3.09**	1.20	0.37	-0.46	-0.54

This table reports t-statistics for the hypothesis that the pricing errors are independent. Bold font is identifies the statistics that are significant at the 5%-significance level. The statistic $t_{(x,60)}$, which compares $RV_{\mathrm{AC}_1}^{(x \text{ ticks})}$ to $RV_{\mathrm{AC}_1}^{(60 \text{ ticks})}$, has power against alternatives for which the dependence endures for more that x ticks.

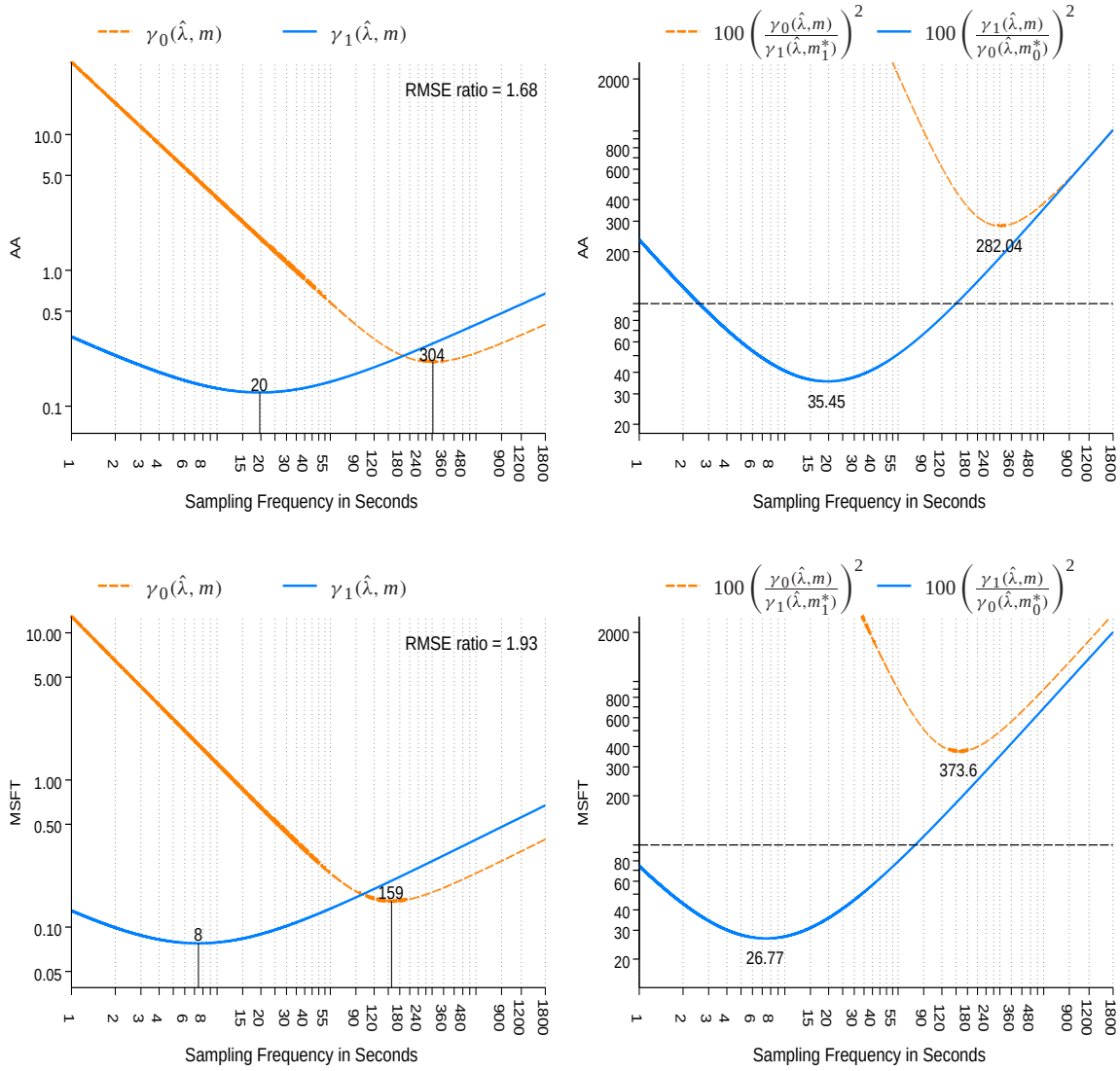

Figure 1: Absolute and relative RMSEs for RV and RV_{AC_1}. The upper panels are results for AA and the lower panels are results for MSFT. The left panels shows $\gamma_0(\hat{\lambda}, m)$ and $\gamma_1(\hat{\lambda}, m)$ using empirical estimates for λ. The right panels show $\gamma_0(\hat{\lambda}, m)/\gamma_1(\hat{\lambda}, m_1^*)$ and $\gamma_1(\hat{\lambda}, m)/\gamma_0(\hat{\lambda}, m_0^*)$ that represent the relative efficiencies of $RV^{(m)}$ and $RV_{\text{AC}_1}^{(m)}$ (relative to $RV_{\text{AC}_1}^{(m_1^*)}$ and $RV^{(m_0^*)}$, respectively). The x-axis refers to $\delta_{i,m} = (b-a)/m$ in units of seconds, where $b - a = 6.5$ hours (a trading day).

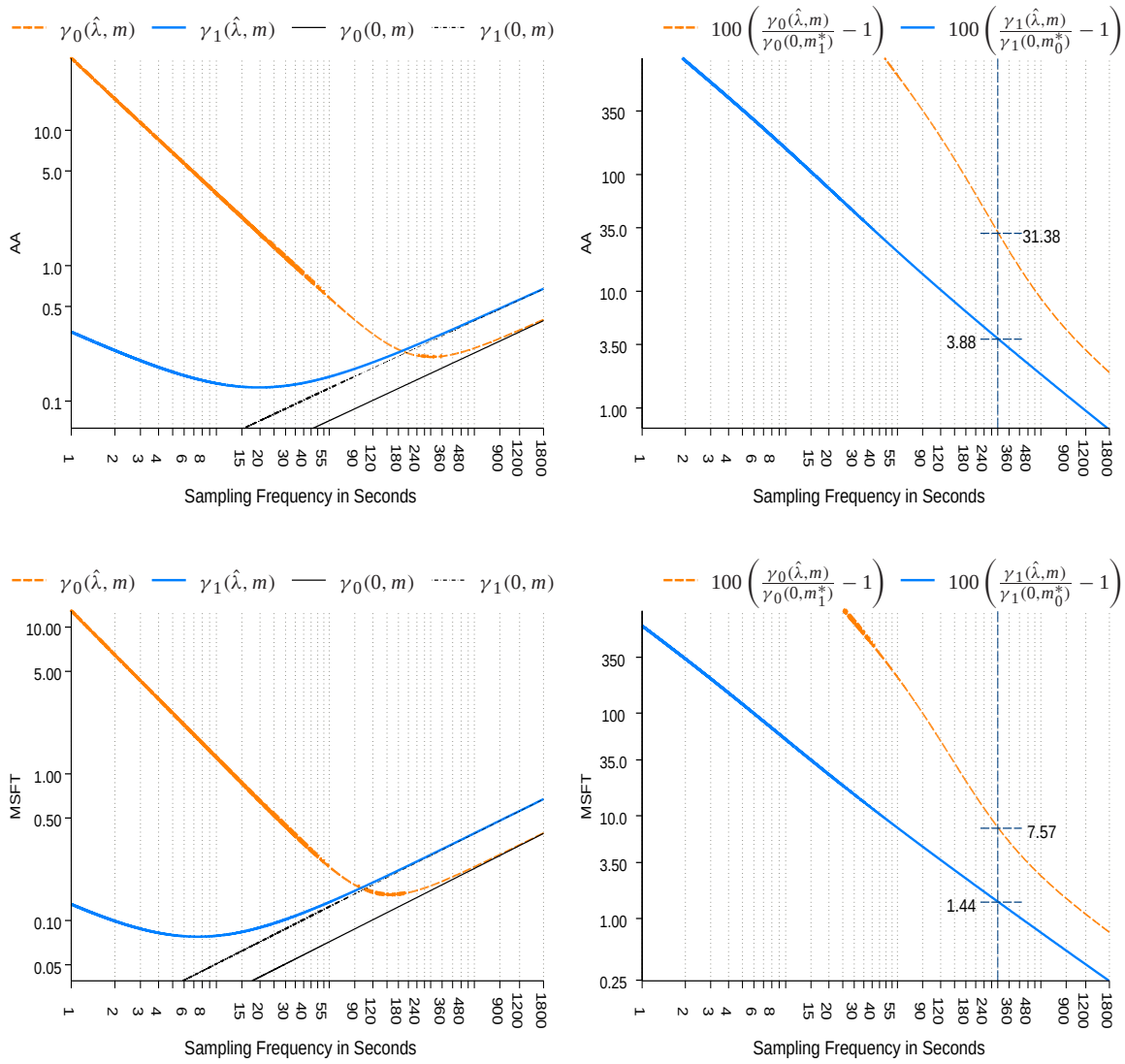

Figure 2: This figure illustrates the effects of market microstructure noise on the RMSE. The left panels contains $\gamma_0(\hat{\lambda}, m)$ and $\gamma_1(\hat{\lambda}, m)$ (wide light lines) and $\gamma_0(0, m)$ and $\gamma_1(0, m)$ (thin dark lines), where the thin dark lines represent the RMSEs of $RV^{(m)}$ and $RV^{(m)}_{\mathrm{AC}_1}$ in the absence of noise ($\omega^2 = 0$). The right panels show the percentage increase of the RMSE that is due to noise as a function of the sampling frequency, m. The x-axis refers to $\delta_{i,m} = (b - a)/m$ in units of seconds, where $b - a = 6.5$ hours (a trading day).

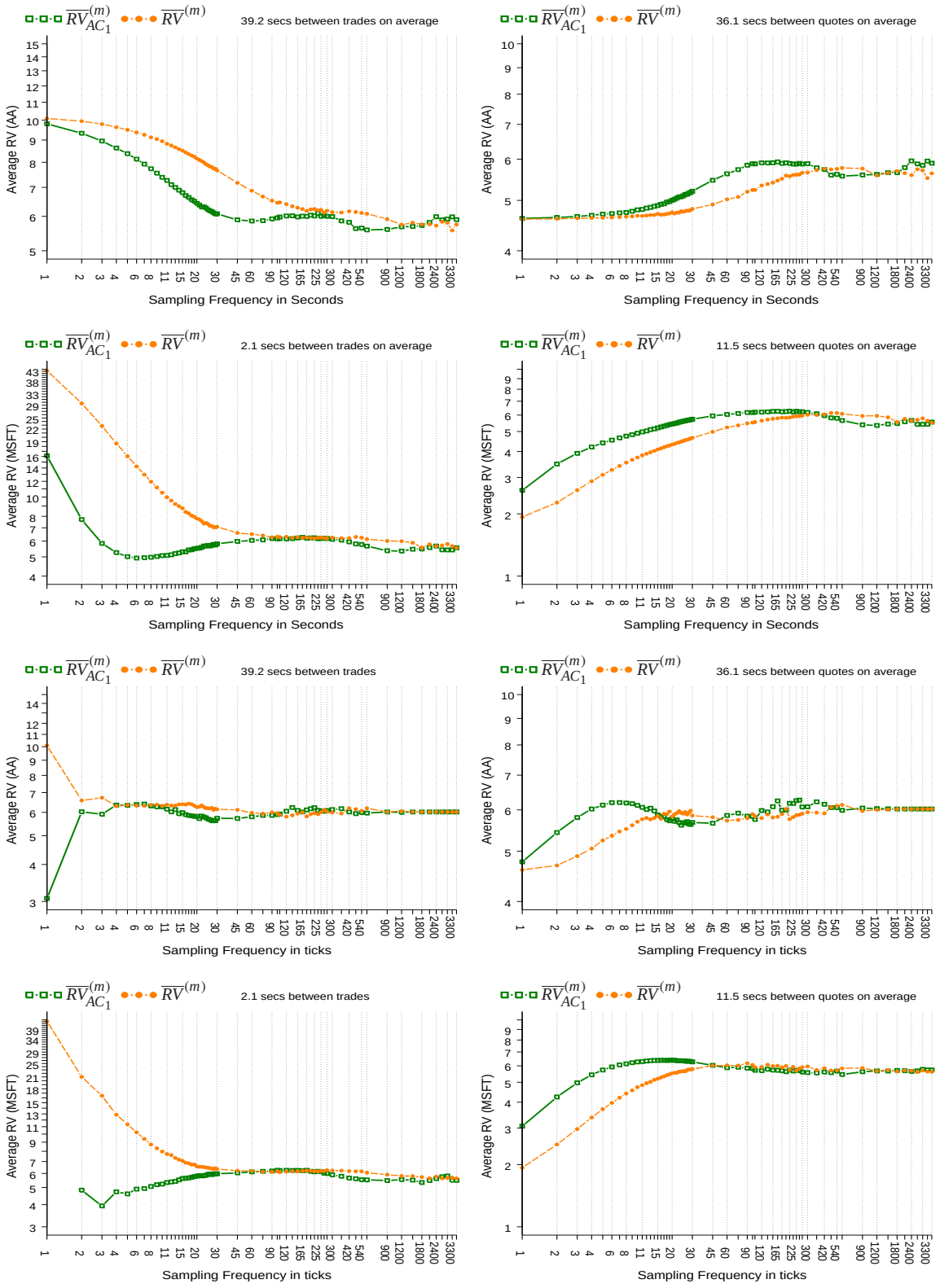

Figure 3: Volatility signature plots for AA and MSFT using transaction data (left panels) and quotation data (right panels). The first two rows of panels are volatility signature plots based on calendar time sampling, whereas the lower two rows are based on transaction time sampling.

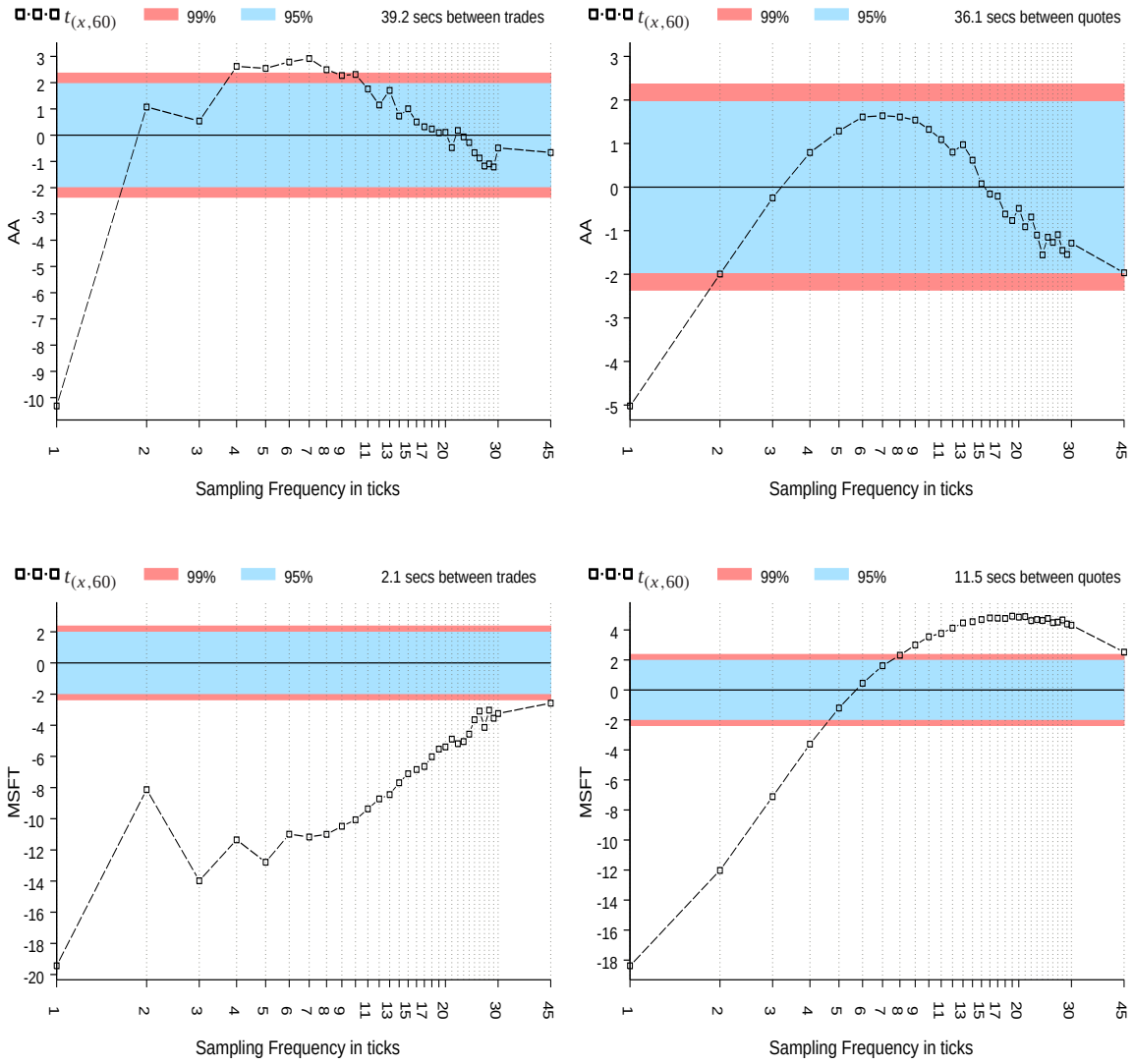

Figure 4: The figure shows the t-statistics, $t_{(x,60)}$, of the hypothesis that the noise process, u, is time-independent. The $t_{(x,60)}$-test has power against alternative for which the time-dependence endures for more than x ticks, so the figure is informative about the time-dependence in u.