Realized Variance and IID Market Microstructure Noise

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

[a] *Brown University, Department of Economics, Box B,Providence, RI 02912, USA*

[b] *Aarhus School of Business, Department of Information Science, Denmark*

Version: February 20, 2004

Abstract

We analyze the properties of a bias-corrected realized variance (*RV*) in the presence of iid market microstructure noise. The bias correction is based on the first-order autocorrelation of intraday returns and we derive the optimal sampling frequency as defined by the mean squared error (MSE) criterion. The bias-corrected *RV* is benchmarked to the standard measure of *RV* and an empirical analysis shows that the former can reduce the MSE by 50%-90%. Our empirical analysis also shows that the iid noise assumption does not hold in practice. While this need not affect the *RV*s that are based on low-frequency intraday returns, it has important implications for those based on high-frequency returns.

Keywords: Realized Variance; High-Frequency Data; Integrated Variance.

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. Andreou & Ghysels (2002) and Oomen (2002*a*).[1] 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 is to use bias correction techniques. In this paper, we analyze an estimator that utilize the first-order autocorrelation to bias-correct the *RV*. This estimator is denoted by RV_{AC_1} and has previously been used by French, Schwert & Stambaugh (1987) and Zhou (1996), who applied it to

*Corresponding author, email: Peter_Hansen@brown.edu

[1] The bias is particularly evident from the so-called *volatility signature plots* that were introduced by Andersen, Bollerslev, Diebold & Labys (2000).

daily returns and intraday returns, respectively.[2] The subscript 'AC_1' refers to the fact that we use one (the first) autocorrelation of intraday returns to correct for the bias.

We make three contributions in this paper. First, we derive the bias and variance properties of the RV_{AC_1} and the optimal sampling frequency as defined by the mean squared error (MSE) criterion. Second, we derive the asymptotic distribution of RV_{AC_1} and show that its asymptotic variance is smaller than that of the standard RV. Third, the analysis is based on a particular type of market microstructure noise, which has previously been analyzed by Corsi, Zumbach, Müller & Dacorogna (2001), Zhang, Mykland & Aït-Sahalia (2003), and Bandi & Russell (2003). Here it is assumed that the noise is independent and identically distributed (across time) and that the noise is independent of the true price process. We label this type of noise as *iid noise*. An important result of our empirical analysis is that the *iid noise* assumption does not hold in practice. Under the *iid noise* assumption the RV_{AC_1} is unbiased at any sampling frequency, however the RV_{AC_1} is clearly biased when returns are sampled at high frequencies. While the RV_{AC_1} should reduce the MSE by 80%–90% compared to the standard RV, when based on its optimal sample frequency (about five-second sampling), we conclude that the implications of the iid noise assumption are only valid when we sample every 30 seconds (or slower). At this sampling frequency the unbiased RV_{AC_1} leads to a reduction of the MSE by a little more than 50% in our empirical analysis.

The paper is organized as follows. In Section 2, we define the RV_{AC_1} and derives its properties. Section 3 contains an empirical analysis that quantifies the relative MSE of RV_{AC_1} to that of the standard RV, and Section 4 contains concluding remarks. All proofs are given 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 process is given by $u(t) \equiv p(t) - p^*(t)$. The noise process, u, may be due market microstructure effects such as bit-ask bounces, but the discrepancy between p and p^* can also be a result of 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 assume that the specification for p^* is a simple stochastic volatility model and our assump-

[2] Other approached to bias correcting the RV include the filtering techniques by Andersen, Bollerslev, Diebold & Ebens (2001) (moving average) and Bollen & Inder (2002) (autoregressive).

[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). Some additional approaches to calculate a measure for the realized variance are discussed in Andersen, Bollerslev & Diebold (2003).

tions about the (continuous-time) noise process, are analogous to standard (discrete-time) assumptions in the literature. We need the following definition.

Definition 1 (Gaussian iid process) *We call $u(t)$ a Gaussian iid process with mean μ and variance ω^2 if $u(t)$ and $u(s)$ are independent for all $t \neq s$ and $u(t) \sim N(\mu, \omega^2)$ for all $t \in \mathbb{R}$.*

Lemma 1 *The Gaussian iid process exists and $(u(t_1), \ldots, u(t_k))' \sim N_k(\boldsymbol{\mu}, \omega^2 \mathbf{I}_k)$ for any k-tuple $(t_1, \ldots, t_k)$ of distinct points, where $\boldsymbol{\mu} = (\mu, \ldots, \mu)$ and $\mathbf{I}_k$ is the $k \times k$ identity matrix.*

Assumption 1 *(i) The true price process is given from $dp^*(t) = \sigma(t)dw(t)$, 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). (ii) The noise process, u, is a Gaussian iid process with mean zero and variance ω^2 that is independent of p^*.*

Although we allow the volatility function, $\sigma(t)$, to be random we shall condition on $\sigma(t)$ in our analysis, because our object of interest is the *integrated variance*, $IV \equiv \int_a^b \sigma^2(t)dt$. 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). This specification for the noise process is similar (or identical) to those in Corsi et al. (2001), Zhang et al. (2003), and Bandi & Russell (2003). Assuming a Gaussian distribution is not crucial but makes the analysis more tractable.

We partition the interval $[a, b]$ into m intervals of equal length, $\Delta_m \equiv (b-a)/m$, and obtain the m returns, $y^*_{i,m} \equiv p^*(a + i\Delta_m) - p^*(a + i\Delta_m - \Delta_m)$, $i = 1, \ldots, m$, that will be referred to as *intraday returns*. Similarly we define $y_{i,m}$ and $e_{i,m}$ to be the increments in p and u, respectively, and note that $e_{i,m} = y_{i,m} - y^*_{i,m}$.

The *realized variance* for p^* is defined by $RV^{(m)}_* \equiv \sum_{i=1}^m y^{*2}_{i,m}$, 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 the ideal estimator it is not a feasible estimator, because p^* is latent. The realized variance of p, which is given by $RV^{(m)} \equiv \sum_{i=1}^m y^2_{i,m}$, is observable but suffers from a well-known bias problem and is inconsistent for the IV.

The bias-variance properties of the $RV^{(m)}$ have been established by Zhang et al. (2003) and Bandi & Russell (2003) under an iid noise assumption. The following lemma summarizes some of their results in our framework, where our Gaussian assumptions lead to more detailed (and simpler) expressions. First we define $\sigma^2_{i,m} \equiv \int_{a+i\Delta_m-\Delta_m}^{a+i\Delta_m} \sigma^2(t)dt$ and we note that $\text{var}(y^*_{i,m}) = E(y^{*2}_{i,m}) = \sigma^2_{i,m}$.

Lemma 2 *Given Assumption 1 it holds that* $E(RV^{(m)}) = IV + 2m\omega^2$, $\text{var}(RV^{(m)}) = 12\omega^4 m + 2\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2 - 4\omega^4 + 2\sum_{i=1}^{m} \sigma_{i,m}^4$, *and the asymptotic distribution is given by*

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

Next, we consider the alternative measure of the realized variance, that is given by

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

This quantity incorporates the empirical first-order autocorrelation which explains the subscript. This modification amounts to a bias reduction that 'works' the same way that robust covariance estimators, such as that of Newey & West (1987), achieve their consistency.

Lemma 3 *Given Assumption 1 it holds that* $E(RV_{\text{AC}_1}^{(m)}) = IV$,

$$\text{var}(RV_{\text{AC}_1}^{(m)}) = 8\omega^4 m + 6\omega^2\left(\sum_{i=1}^{m} \sigma_{i,m}^2 - \omega^2\right) + 6\sum_{i=1}^{m} \sigma_{i,m}^4 + \omega^2(\sigma_{0,m}^2 + \sigma_{m+1,m}^2) + O(m^{-2}),$$

and the asymptotic distribution is given by

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

An important result of Lemma 3 is that $RV_{\text{AC}_1}^{(m)}$ is unbiased for the *IV* (conditionally on $\{\sigma(s), a \le s \le b\}$), such that an unbiased measure is available in the presence of market microstructure noise. A rather remarkable result of Lemma 3 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 leads to a larger asymptotic variance. Also note that the asymptotic results of Lemma 3 is more useful than that of Lemma 2, because the result of Lemma 2 does not involve the object of interest, *IV*, but only shed light on aspects of the *RV*'s bias. Note, however, that the asymptotic result of Lemma 3 does not suggest that $RV_{\text{AC}_1}^{(m)}$ should be sampled at the highest possible frequency, since the asymptotic variance is increasing in m. Our expression for the variance is approximately given by $\text{var}[RV_{\text{AC}_1}^{(m)}] \approx 8\omega^4 m + 6\omega^2[\int_a^b \sigma^2(s) - \omega^2] + 6\int_a^b \sigma^4(s) ds \frac{1}{m}$, where the last term involves the *integrated quarticity* that was introduced by Barndorff-Nielsen & Shephard (2002).

Next we compare $RV_{\text{AC}_1}^{(m)}$ to $RV^{(m)}$ in terms of their mean square error (MSE) and their respective optimal sampling frequencies for a special case.

Corollary 4 *Suppose that the volatility is constant such that* $\sigma_{i,m}^2 = \sigma^2/m$, *where* $\sigma^2 = IV$ *and define the noise-to-signal ratio,* $\lambda \equiv \omega^2/\sigma^2$. *The mean squared errors are given by*

$$\text{MSE}(RV^{(m)}) \quad = \quad 2\sigma^4[2\lambda^2 m^2 + 6\lambda^2 m + (\lambda - 2\lambda^2) + \frac{1}{m}],$$

$$\text{MSE}[RV^{(m)}_{\text{AC}_1}] \;=\; 2\sigma^4[\qquad\qquad 4\lambda^2 m + 3(\lambda-\lambda^2) + \frac{3+\lambda}{m}].$$

Let m_0^* and m_1^* be the optimal sampling frequencies for $RV^{(m)}$ and $RV^{(m)}_{\text{AC}_1}$, respectively. It holds that m_0^* is given implicitly as the real (positive) solution to $2m^3+3m^2 = 1/(2\lambda^2)$ whereas $m_1^* = \sqrt{3+\lambda}/(2\lambda)$.

It can be verified that m_1^* is several times larger than m_0^*, thus the optimal $RV^{(m)}_{\text{AC}_1}$ requires more frequent sampling that the 'optimal' RV. This is quite intuitive, because $RV^{(m)}_{\text{AC}_1}$ can utilized more information in the data without being affected by a severe bias.

3. Empirical Analysis

We analyze the Alcoa Inc. (AA) stock over a sample period that spans the five year from January 2, 1998 to December 31, 2002. The data are transaction prices from the NYSE extracted from the Trade and Quote (TAQ) database. The raw data were filtered for outliers and we discarded transactions outside 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 by 13 days. Thus we used the previous-tick method to construct the RVs for a total of $n = 1,242$ days and denoted these by $RV^{(m)}_t$ and $RV^{(m)}_{\text{AC}_1,t}$, $t = 1,\ldots,n$. The RVs are calculated for the hours that the market is open, approximately 390 minutes per day (6.5 hours) for most days.

From Lemmas 2 and 3 it follows that $2m\omega^2 = E[RV^{(m)} - RV^{(m)}_{\text{AC}_1}]$ such that $\hat{\omega}^2 = \frac{1}{2m}(\overline{RV}^{(m)} - \overline{RV}^{(m)}_{\text{AC}_1})$ is a natural estimator of ω^2 (under the assumptions of Corollary 4), where we define the sample averages, $\overline{RV}^{(m)} \equiv n^{-1}\sum_{t=1}^n RV^{(m)}_t$ and $\overline{RV}^{(m)}_{\text{AC}_1} \equiv n^{-1}\sum_{t=1}^n RV^{(m)}_{\text{AC}_1,t}$. With $m = 390$ (1-minute intraday returns) we find that $\overline{RV}^{(m)} - \overline{RV}^{(m)}_{\text{AC}_1} = 0.657$ which leads to $\hat{\omega}^2 = 0.657/(2*390) = 0.000842$, and since $\overline{RV}^{(m)}_{\text{AC}_1} = 4.762$ we obtain $\hat{\lambda} = 0.000842/4.762 = 0.000177$. This leads to $m_0^* \approx 200$ and $m_1^* \approx 4,890$, which corresponds to intraday returns that are sampled approximately every 2 minutes and every 5 seconds, respectively.[4] By plugging these numbers into the formulae of Corollary 4 we find the relative mean squared error to be $\text{MSE}(RV^{(m_0^*)})/\text{MSE}(RV^{(m_1^*)}_{\text{AC}_1}) \approx 4.88$, which (in theory) implies that $RV^{(m_1^*)}_{\text{AC}_1}$ is almost five time more efficient than $RV^{(m_0^*)}$ in terms of the mean squared error criterion. The most commonly used sampling frequency is 5-minute sampling, which corresponds to $m = 78$ in our application. As noted by Bandi & Russell (2003) this results in an additional loss of efficiency and theoretically we have that $\text{MSE}(RV^{(78)})$ is about 10 times larger than $\text{MSE}(RV^{(m_1^*)}_{\text{AC}_1})$.

[4] Bandi & Russell (2003) reported optimal sample frequencies for $RV^{(m)}$ (for several assets) that are quite similar to our estimate of m_0^*.

From Corollary 4 we observe that the root mean squared errors are proportional to σ^2, such that $\text{RMSE}(RV^{(m)}) = \sigma^2 c_{RV}(m)$ and $\text{RMSE}(RV^{(m)}_{\text{AC}_1}) = \sigma^2 c_{AC}(m)$ where $c^2_{RV}(m) \equiv 2[2\lambda^2 m^2 + 6\lambda^2 m + (\lambda - 2\lambda^2) + \frac{1}{m}]$ and $c^2_{AC}(m) \equiv 2[4\lambda^2 m + 3(\lambda - \lambda^2) + \frac{3+\lambda}{m}]$. In the left panel of Figure 1 we have plotted $c_{RV}(m)$ and $c_{AC}(m)$ using our empirical estimate of λ. This reveals that the $RV^{(m)}_{\text{AC}_1}$ dominate the $RV^{(m)}$ except at the lowest frequencies. The left panel also shows that the $RV^{(m)}_{\text{AC}_1}$ is less sensitive to the choice of m. This is also clear from the right panel of Figure 1, where we have displayed the relative MSE of $RV^{(m)}_{\text{AC}_1}$ to that of (the optimal) $RV^{(m^*_0)}$ and the relative MSE of $RV^{(m)}$ to that of (the optimal) $RV^{(m^*_1)}_{\text{AC}_1}$. One aspect that can be read of Figure 1 is that the $RV^{(m)}_{\text{AC}_1}$ 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.

[Figure 1 about here]

The optimal sample frequencies of Corollary 4 depend on parameters that are likely to differ across days. So our estimates above should be viewed as approximations for 'daily average values', in the sense that $m_0 = 200$ is a sensible sampling frequency to use (on average), although different values are likely to be better on some days. While m^*_1 indicate that we should sample intraday returns every 5 seconds, we shall see that the implications of the iid noise assumption do not hold in practice if intraday returns are sampled at high frequencies. In our application the implications seem to fail once intraday returns are sampled more frequently than every 30 seconds.

3.1. *Empirical Evidence against the IID Noise Assumption*

Under the iid noise assumption the $RV^{(m)}_{\text{AC}_1}$ should be unbiased at any frequency. This can be understood from the fact that the iid noise assumption causes the first-order autocorrelation of $e_{i,m}$ (and hence $y_{i,m}$) to be non-zero, whereas higher-order covariances are all zero. The $RV^{(m)}_{\text{AC}_1}$ properly corrects for the first-order autocorrelation in $y_{i,m}$, which is the reason that the $RV^{(m)}_{\text{AC}_1}$ is unbiased under the iid assumption. If higher-order autocorrelations of $y_{i,m}$ are non-zero, which could be the case if the noise component, $u(t)$, was dependent across time (different from iid noise), then the $RV^{(m)}_{\text{AC}_1}$ would be biased (for large ms). This problem is evident from the signature plots in Figure 2 that show that the $RV^{(m)}_{\text{AC}_1}$ is biased for sampling frequency above 30 seconds. For example, with 1-second sampling the bias is quite severe and close to that of the standard RV, however the $RV^{(m)}_{\text{AC}_1}$ generally has a smaller bias.

[Figure 2 about here]

In spite of this shortcoming, we will still argue that the $RV^{(m)}_{\mathrm{AC}_1}$ is preferred to the standard RV. The volatility signature plot of RV_{AC_1} indicate that the time-dependence in u persists for less than 30 seconds, because the signature plot is quite constant for the frequencies that are below a 30-second sampling. So our estimate of λ (that is based on 1-minute returns) should not be affected by the time dependence, and this value of λ suggests that the MSE of the $RV^{(780)}_{\mathrm{AC}_1}$ (30-seconds returns) is 58% smaller than that of the 'optimal' $RV^{(m^*_0)}$, see Figure 1. Nevertheless, Figure 2 shows that there is a need to study the properties of the RV under a more general specification for the noise process, such as the Ornstein–Uhlenbeck specification that was analyzed in a related setting by Aït-Sahalia, Mykland & Zhang (2003).

4. Concluding Remarks

We have derived the bias and variance properties of $RV^{(m)}_{\mathrm{AC}_1}$, which equals the standard realized variance plus a bias correction that is given from the first-order autocorrelation of intraday returns. The $RV^{(m)}_{\mathrm{AC}_1}$ compares favorable to the standard measure of RV in terms of the mean squared error criterion. Our empirical analysis showed that the MSE of $RV^{(m)}_{\mathrm{AC}_1}$ may be 90% smaller than the MSE of the most common measure of RV, provided that the market microstructure noise satisfies the iid assumption. Most of the existing theoretical studies of the RV in the presence of market microstructure effects are based on this assumption, however our empirical analysis revealed that this assumption does not hold in practice. While it may be true (or approximately true) for sampling at low frequencies, it does not hold when returns are sampled more frequently than every 30 seconds in our empirical analysis. This followed directly from the volatility signature plot of $RV^{(m)}_{\mathrm{AC}_1}$ in Figure 2. While the $RV^{(m)}_{\mathrm{AC}_1}$ is biased when sampling at high frequencies, its bias was less severe than that of the standard RV, and $RV^{(m)}_{\mathrm{AC}_1}$ was found to dominate the standard $RV^{(m)}$ when the former is based on a less aggressive sampling, such as 30-second sampling. However, our analysis has revealed a need to study the properties of RV-measures under a more general specification for the noise process. Some preliminary results can be found in Hansen & Lunde (2003) who use a model-free noise structure, and in Oomen (2002*b*) who use a model-based approach.

Acknowledgements

We thank Neil Shephard for valuable comments. Financial support from the Danish Research Agency, grant no. 24-00-0363 is gratefully acknowledged. All errors remain our responsibility.

Appendix of Proofs

Proof of Lemma 1. That $(u(t_1), \ldots, u(t_k))' \sim N(\mu, \omega^2 I_k)$ follows from the definition of u, and since this is a well-defined (multivariate) Gaussian distribution, the existence of u follows directly from Kolmogorov's Existence Theorem, see Billingsley (1995, chapter 7). ■

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

Proof of Lemma 2. 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 $E(e_{i,m}^2) = 2\omega^2$. 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

$$\begin{aligned}
E(e_{i,m}^4) &= E(u_{i,m} - u_{i-1,m})^4 = E(u_{i,m}^2 + u_{i-1,m}^2 - 2u_{i,m}u_{i-1,m})^2 \\
&= E(u_{i,m}^4 + u_{i-1,m}^4 + 4u_{i,m}^2 u_{i-1,m}^2 + 2u_{i,m}^2 u_{i-1,m}^2) + 0 \\
&= 6\omega^4 + 6\omega^4 = 12\omega^4, \\
E(e_{i,m}^2 e_{i+1,m}^2) &= E(u_{i,m} - u_{i-1,m})^2 (u_{i+1,m} - u_{i,m})^2 \\
&= E(u_{i,m}^2 + u_{i-1,m}^2 - 2u_{i,m}u_{i-1,m})(u_{i+1,m}^2 + u_{i,m}^2 - 2u_{i+1,m}u_{i,m}) \\
&= E(u_{i,m}^2 + u_{i-1,m}^2)(u_{i+1,m}^2 + u_{i,m}^2) + 0 = 6\omega^4.
\end{aligned}$$

such that $\text{var}(e_{i,m}^2) = 12\omega^4 - [E(e_{i,m}^2)]^2 = 8\omega^4$ and $\text{cov}(e_{i,m}^2, e_{i+1,m}^2) = 2\omega^4$. Since $\text{cov}(e_{i,m}^2, e_{i+h,m}^2) = 0$ for $|h| \geq 2$ it follows that

$$\text{var}(\sum_{i=1}^m e_{i,m}^2) = \sum_{i=1}^m \text{var}(e_{i,m}^2) + \sum_{\substack{i,j=1 \\ i \neq j}}^m \text{cov}(e_{i,m}^2, e_{i+h,m}^2) = m8\omega^4 + 2(m-1)2\omega^4 = 12m\,\omega^4 - 4\omega^4.$$

The last sum involves uncorrelated terms such that

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

Finally, the asymptotic normality follows by the central limit theorem for heterogeneous arrays with finite dependence, and the fact that $2\sum_{i=1}^m \sigma_{i,m}^4 + 2\omega^2 \sum_{i=1}^m \sigma_{i,m}^2 - 4\omega^4 = O(1)$. ■

Proof of Lemma 3. First we note that $RV_{\text{AC}_1}^{(m)} = \sum_{i=1}^m Y_{i,m} + U_{i,m} + V_{i,m} + W_{i,m}$, 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 $W_{i,m}$. 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 $E[RV^{(m)}_{\mathrm{AC}_1}] = \sum_{i=1}^m \sigma^2_{i,m}$, and 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 $(1) = \mathrm{var}(\sum_{i=1}^m Y_{i,m})$, $(2) = \mathrm{var}(\sum_{i=1}^m U_{i,m})$, $(3) = \mathrm{var}(\sum_{i=1}^m V_{i,m})$, $(4) = \mathrm{var}(\sum_{i=1}^m W_{i,m})$, $(5) = \mathrm{cov}(\sum_{i=1}^m V_{i,m}, \sum_{i=1}^m W_{i,m})$, since all other sums are uncorrelated. Next, we derive 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 our assumptions 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

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

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

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

$$= \quad E[u_{i,m}u_{i+1,m}u_{i+1,m}u_{i,m}] + 0 = \omega^4,$$

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 $\text{var}(V^2_{i,m}) = \sigma^2_{i,m}2\omega^2$ and $E[V_{i,m}V_{i+h,m}] = 0$ for all $h \neq 0$. Thus $(3) = \text{var}(\sum_{i=1}^m V_{i,m}) = 2\omega^2 \sum_{i=1}^m \sigma^2_{i,m}$.

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

$$\text{cov}(W_{i,m}, W_{i+1,m}) = E[-u^2_{i,m}(y^{*2}_{i,m} + y^{*2}_{i+1,m})] = -\omega^2(\sigma^2_{i,m} + \sigma^2_{i+1,m}),$$

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

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

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

By adding up the five terms we find

$$\begin{aligned}
&6\sum_{i=1}^m \sigma^4_{i,m} - 2\sum_{i=1}^{m-1}(\sigma^2_{i+1,m} - \sigma^2_{i,m})^2 - 2(\sigma^4_{1,m} + \sigma^4_{m,m}) + \sigma^2_{1,m}\sigma^2_{0,m} + \sigma^2_{m,m}\sigma^2_{m+1,m} + 8\omega^4 m - 6\omega^4 \\
&+ 2\omega^2 \sum_{i=1}^m \sigma^2_{i,m} + 2\omega^2 \sum_{i=1}^m \sigma^2_{i,m} + 2\omega^2[\sigma^2_{0,m} + \sigma^2_{m+1,m}] + 2\sum_{i=1}^m \sigma^2_{i,m}\omega^2 - \omega^2[\sigma^2_{1,m} + \sigma^2_{m,m}] \\
&= 8\omega^4 m + 6\omega^2(\sum_{i=1}^m \sigma^2_{i,m} - \omega^2) + 6\sum_{i=1}^m \sigma^4_{i,m} + \omega^2(\sigma^2_{0,m} + \sigma^2_{m+1,m}) + \kappa_m,
\end{aligned}$$

where

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

Since $\sigma^2(t)$ is Lipschitz, there exists an $\epsilon > 0$ such that $|\sigma^2(t) - \sigma^2(t+\delta)| \leq \epsilon\delta$ for all t and all δ. Thus if we define the interval, $J_{i,m} \equiv [a + (i-1)\Delta_m, a + i\Delta_m]$, we have that $|\sigma^2_{i,m}| = |\int_{J_{i,m}} \sigma^2(s)ds| \leq \Delta_m \sup_{s\in J_{i,m}} \sigma^2(s) = O(m^{-1})$, since $\Delta_m = (b-a)/m = O(m^{-1})$, and

$$|\sigma^2_{i,m} - \sigma^2_{i-1,m}| \quad = \quad |\int_{J_{i,m}} \sigma^2(s) - \sigma^2(s - \Delta_m)ds| \leq \int_{J_{i,m}} |\sigma^2(s) - \sigma^2(s - \Delta_m)|ds$$

$$\leq \quad \Delta_m \sup_{s \in J_{i,m}} |\sigma^2(s) - \sigma^2(s - \Delta_m)| \leq \Delta_m^2 \epsilon = O(m^{-2}).$$

Finally, $\sum_{i=1}^{m} (\sigma^2_{i+1,m} - \sigma^2_{i,m})^2 \leq m \cdot (\Delta_m \frac{\epsilon}{m})^2 = O(m^{-3})$, which proves that $\kappa_m = O(m^{-2})$. The asymptotic normality follows from the CLT that applies to heterogeneous arrays with finite dependence, since $y_{i,m}(y_{i-1,m} + y_{i,m} + y_{i+1,m})$ is a finite dependent (3-dependent) process for any m. ■

Proof of Corollary 4. The MSE's are given from Lemmas 2 and 3. Setting the $\partial \text{MSE}(RV^{(m)})/\partial m \propto 4\lambda^2 m + 6\lambda^2 - m^{-2}$ equal to zero yields the first order condition of the corollary. Similarly we find $\partial \text{MSE}(RV^{(m)}_{\text{AC}_1})/\partial m \propto 4\lambda^2 - (3+\lambda)m^{-2}$, which proves that $m_1^* = \sqrt{3+\lambda}/(2\lambda)$. ■

References

Aït-Sahalia, Y., Mykland, P. A. & Zhang, L. (2003), How often to sample a continuous-time process in the presence of market microstructure noise, Working Paper w9611, NBER.

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. (2003), Parametric and nonparametric volatility measurement, *in* Y. Aït-Sahalia & L. P. Hansen, eds, 'forthcoming in Handbook of Financial Econometrics', Vol. I, Elsevier-North Holland, Amsterdam.

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.

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.

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.

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

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.

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

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*a*), 'Modelling realized variance when returns are serially correlated'. manuscript, Warwick Business School, The University of Warwick.

Oomen, R. A. C. (2002*b*), 'Statistical models for high frequency security prices'. manuscript, Warwick Business School, The University of Warwick.

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. (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.

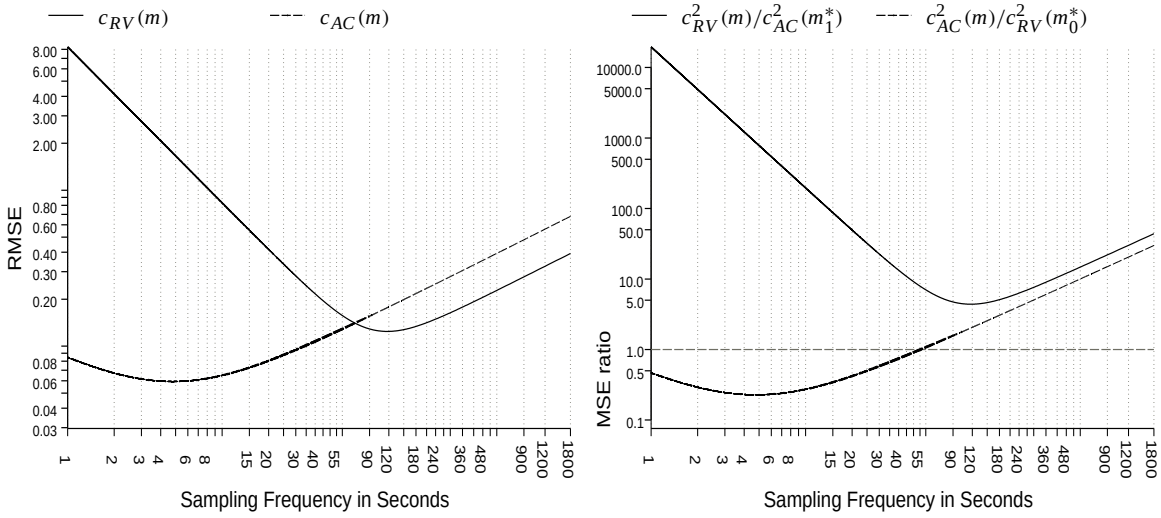

Figure 1: Left: The MSEs of $RV^{(m)}$ and $RV^{(m)}_{AC_1}$ as a function of the sampling frequency, m. Right: Relative MSE of $RV^{(m)}$ to $RV^{(m^*_1)}_{AC_1}$ where m^*_1 is the optimal sampling frequency for RV_{AC_1}, and relative MSE of $RV^{(m)}_{AC_1}$ to $RV^{(m^*_0)}$ where m^*_0 is the optimal sampling frequency for the standard RV.

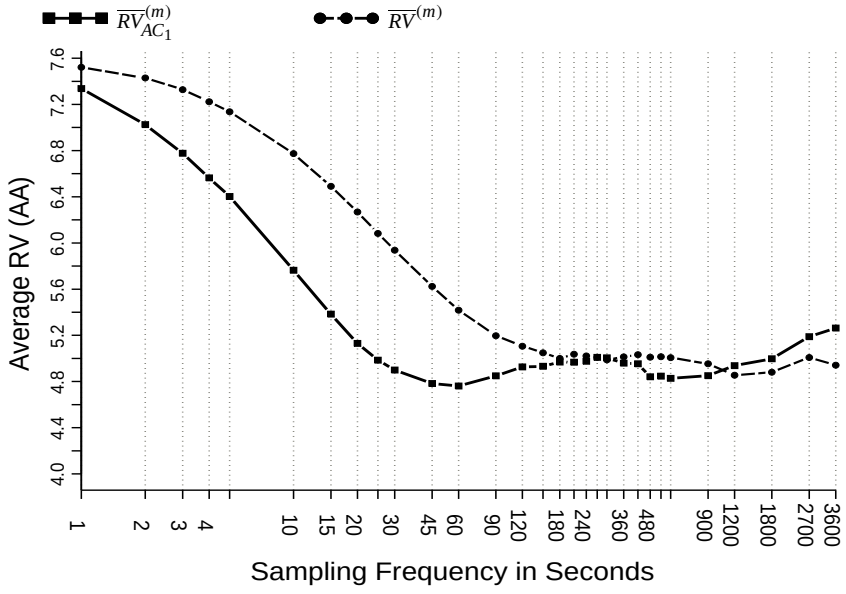

Figure 2: Signature plots of the standard $RV^{(m)}$ and the bias corrected $RV^{(m)}_{AC_1}$.