Properties of Bias Corrected Realized Variance in Calendar Time and Business Time

Roel C.A. Oomen[*]

Department of Accounting and Finance

Warwick Business School

The University of Warwick

Coventry CV4 7AL, United Kingdom

E-mail: roel.oomen@wbs.ac.uk

April 2004

Abstract

In this paper I study the statistical properties of a bias corrected realized variance measure when high frequency asset prices are contaminated with market microstructure noise. The analysis is based on a pure jump process for asset prices and explicitly distinguishes among different sampling schemes. Two important findings emerge from the theoretical and empirical analysis. Firstly, business time sampling is generally superior to the common practice of calendar time sampling in terms of MSE of realized variance. Secondly, a first order bias correction to realized variance allows for the efficient use of much higher frequency data and results in substantial improvements of the statistical properties of realized variance. For IBM transaction data, I estimate a 12 second optimal sampling frequency and a drop of more than 65% in bias and MSE when a first order bias correction is applied.

Keywords: Market Microstructure Noise; Bias Corrected Realized Variance; Optimal Sampling

JEL Classifications: G12, C13, C14, C22

1 Introduction

An all important issue which stands in between the theory and practice of realized variance, is the emergence of a market microstructure induced contamination of prices at high sampling frequencies. In some sense the challenge is to balance the tension between the strong desire for a model-free volatility estimator that efficiently exploits the information contained in high frequency data on the one hand, and the statistical complications that are introduced by the wide range of existing market microstructure effects on the other. This paper aims to contribute to this active debate by providing an in-depth analysis of a modified realized variance measure that explicitly account for market microstructure effects and will be shown to yield accurate and reliable volatility estimates, even when prices are sampled at extremely high frequencies. In particular, I study a bias corrected realized variance that can be traced back to the work by French, Schwert, and Stambaugh (1987) and has recently been revived by Hansen and Lunde (2004b). Different from the existing literature, I study the estimator in the context of a pure jump process for asset prices and analyze the impact that a particular choice of sampling scheme has on the statistical properties of the (bias corrected) realized variance. The main finding of the paper is that, in the presence of market microstructure noise, a simple first order bias correction leads to a substantial improvement of the statistical properties of realized variance. Sampling the price process in business time, as opposed to the common practice of calendar time sampling, further improves matters.

To model high frequency security prices, I adopt a generalization of the Press (1967) compound Poisson process that has recently been proposed by Oomen (2004). In particular, I assume that asset prices follow an heterogeneous compound Poisson process with serially correlated innovations. Since the joint characteristic function of returns is available, closed form expressions for the bias and mean squared error of the bias corrected realized variance can be derived. The pure jump process is fundamentally different from the diffusion based ones considered in closely related work by Bandi and Russell (2003), Hansen and Lunde (2004a), and Zhang, Mykland, and Ait-Sahalia (2003), and the theoretical results derived in this paper should therefore been seen to complement existing ones. Based on this model, the paper provides a framework which can be used in practice to (i) determine the optimal sampling frequency of a price process on a daily basis, (ii) measure the benefits of sampling in business time, and (iii) measure the benefits of bias correcting the realized variance. Using IBM transaction data, I estimate an optimal sampling frequency for realized variance of about 2.5 minutes. When the proposed bias correction is applied, realized variance can be computed based on prices sampled at frequencies as high as 12 seconds with an associated reduction in mean squared error of more than 65%. Interestingly, these results are directly in line with those reported in Bandi and Russell (2003) and Hansen and Lunde (2004a) which both use very different methodologies.

The remainder of this paper is structured as follows. Section 2 introduces the bias corrected realized variance and reviews some of the key features of the jump model for high frequency prices on which the analysis is based. This is then followed by a discussion of the statistical properties of the bias corrected realized variance. Section 3 reports results for IBM and S&P 500 Spider transaction data while section 4 concludes.

2 Bias Corrected Realized Variance under Alternative Sampling Schemes

Let $\{P(t), t \in [0, T]\}$ denote the logarithmic asset price process. Without loss of generality, we focus on the unit time interval, i.e. $T = 1$, which in practice typically spans one trading day. Given a discrete set of price observations, $\{P(t_i)\}_{i=0}^{N}$ where $t_0 = 0$, $t_N = 1$, and $t_i < t_{i+1}$, realized variance is defined as:

$$RV_N = \sum_{i=1}^{N} R(t_i | \tau_i)^2,$$

where $R(t|\tau) \equiv P(t) - P(t - \tau)$ and $\tau_i = t_i - t_{i-1}$. It is well known that when P is a semi-martingale, realized variance is a consistent (for $N \to \infty$) estimator of the quadratic variation (QV) of the price process (e.g. Protter (1990)). Because QV is not only a natural measure of volatility that is closely related, and in some cases equal to the conditional return variance (e.g. Andersen, Bollerslev, Diebold, and Labys (2003)), but is also a key input for option pricing and risk management problems, the motivation for the use of realized variance is clear. Moreover, Barndorff-Nielsen and Shephard (2004) have established that the asymptotic distribution of realized variance is Gaussian for a wide range of stochastic volatility models relevant to finance, thereby further adding to its appeal.

Despite the generality and elegance of the theory of realized variance, its practical implementation can be fraught with complications. First of all, N cannot be chosen freely but is dictated by the frequency at which the data is available. While it is not uncommon to observe several thousand prices per day for liquid securities such as IBM, or even larger amounts of data for some exchange rate related securities, the majority of asset prices are observed at much lower frequency (Easley, Kiefer, O'Hara, and Paperman 1996). In such cases, when N is relatively small, realized variance may be very noisy, the asymptotic distribution may not provide a good approximation to the finite sample distribution, and a model based volatility estimator may well yield superior results. On the other hand, when N is large, prices at the highest frequency are often severely contaminated by market microstructure effects which induce serial correlation in returns and render realized variance biased. Early recognition of this issue is provided by Roll (1984), who shows that the bid-ask bounce leads to negative serial correlation in returns, and French, Schwert, and Stambaugh (1987) who emphasize that non-synchronous trading in indices leads to positive serial correlation and biased volatility estimates. Indeed, Froot and Perold (1995) find that for the S&P500 cash index from 1983-1989, volatility estimates based on weekly data are significantly higher (about 20%) than those based on 15-minute data. More recently, a large number of papers have documented severe market microstructure induced biases in realized variance for a wide range of securities, time horizons, and sampling frequencies (e.g. Andersen, Bollerslev, Diebold, and Labys (2000), Bandi and Russell (2003), Corsi, Zumbach, Müller, and Dacorogna (2001), Hansen and Lunde (2004a), Oomen (2003, 2004)).

To deal with the consequences of market microstructure on realized variance, several approaches can be taken. For example, one can aggregate returns down to a frequency at which market microstructure effects have faded and the bias of realized variance is eliminated, or, sample the price process at a frequency that minimizes the mean squared error or some other criterion function of interest. The determination of such an "optimal sampling frequency" has been studied by Bandi and Russell (2003) and Oomen (2004). Alternatively, one can filter the

contaminated price data to purge out any serial correlation (Andersen, Bollerslev, Diebold, and Ebens 2001) or explicitly model the market microstructure effects and use this to bias correct the realized variance (Corsi, Zumbach, Müller, and Dacorogna 2001). When the inherent loss of information associated with aggregation, and the need to specify a possibly misspecified model is to be avoided, Zhang, Mykland, and Ait-Sahalia (2003) propose an approach which involves the averaging of realized variances over sub-samples of the data while Hansen and Lunde (2004a, b) study a bias corrected realized variance measure. The latter approach, which is inspired by Newey and West (1987) and has previously been implemented by French, Schwert, and Stambaugh (1987) and Zhou (1996), is also the focuss of this paper. In particular, I study the properties of the following bias corrected realized variance:

$$RVAC_{N,q} = \sum_{i=1}^{N} R\left(t_i|\tau_i\right)^2 + \sum_{i=1}^{N} R\left(t_i|\tau_i\right) \sum_{k=1}^{q} \left(R\left(t_{i-k}|\tau_{i-k}\right) + R\left(t_{i+k}|\tau_{i+k}\right)\right). \tag{1}$$

The quadratic variation of the price process is thus estimated as the sum of squared returns plus a correction term that is based on the first q empirical autocovariances. Both the theoretical and empirical analysis below illustrate the effectiveness of this estimator in that it dramatically reduces the bias and MSE of realized variance in the presence of market microstructure induced serial correlation of returns. Even though these results are directly in line with the closely related work by Hansen and Lunde (2004a, b), this paper stands to make several contributions to the existing literature. Firstly, the bias corrected realized variance is studied in the context of a *pure jump* process for high frequency security prices. So far, most of the work on realized variance has exclusively focussed on diffusion-based models and the results derived in this paper therefore complement existing ones and lead to new insights into the properties of realized variance. Secondly, the bias corrected realized variance is studied under *alternative sampling schemes*. While it is common practice to sample prices at regularly spaced intervals in calendar time, neither theory nor practice prevents the use of irregularly spaced data. The results in this paper indicate that the scheme used to sample high frequency prices can have an important impact on the statistical properties of (bias corrected) realized variance. Importantly, I find that sampling regularly spaced in business time is generally superior to sampling regularly spaced in calendar time in that it leads to a lower MSE of realized variance. To the best of my knowledge, this issue has not been considered in the literature so far.

2.1   A Jump Process for High Frequency Security Prices

In this paper I model the logarithmic asset price as a heterogeneous compound Poisson process with serially correlated increments. The model is identical to the one considered by Oomen (2004), but to establish notation and facilitate presentation I include a brief review and outline some of its key features relevant to the analysis in this paper. Let the logarithmic asset price, $P\left(t\right)$, be given by:

$$P\left(t\right) = P\left(0\right) + \sum_{j=1}^{M(t)} \left(\varepsilon_j + \eta_j\right) \qquad \text{where} \qquad \eta_j = \rho_0 \nu_j + \rho_1 \nu_{j-1} + \ldots + \rho_q \nu_{j-s}, \tag{2}$$

where $\varepsilon_j \sim \mathcal{N}\left(\mu_\varepsilon, \sigma_\varepsilon^2\right)$ and[1] $\nu_j \sim \mathcal{N}\left(\mu_\nu, \sigma_\nu^2\right)$ and $M(t)$ is a Poisson process with instantaneous intensity $\lambda(t) > 0$. Since the focuss here is on the analysis of financial transaction data, I interpret and refer to $\lambda(t)$ as the instantaneous arrival frequency of trades with the process $M(t)$ counting the number of trades that have occurred up to time t. As such, the model is closely related to the literature on subordinated processed initiated by Clark (1973). Viewing the price process at the kth transaction, i.e.

$$P_k = P_0 + \sum_{j=1}^{k} \varepsilon_j + \sum_{j=1}^{k} \eta_j = P_k^e + \sum_{j=1}^{k} \eta_j,$$

suggests the following interpretation: the observed but market microstructure contaminated price, P, is equal to the efficient but unobserved price, P^e, plus a noise component. To prevent accumulation of noise under temporal aggregation the MA(s) parameters are restricted such that $\eta_k = \sum_{i=0}^{s-1} \bar{\rho}_j \left(\nu_{k-i} - \nu_{k-i-1}\right)$ and the resulting model is referred to as the "restricted CPP-MA(s)" model throughout the remainder of this paper. To further streamline exposition, I will assume that $\mu_\varepsilon = \mu_v = 0$ and concentrate on the simplest case where the order of the MA(s) process is 1 and $\rho_0 = -\rho_1 = 1$. Even though it is straightforward to derive results in full generality, it needlessly complicates notation and obscures the main point of the paper.

The restricted CPP-MA(1) model specified above has two key features that are crucial for the analysis in this paper. First, the MA dependence on the noise component, η, serves to capture the serial correlation in returns induced by market microstructure effects and allows for an analysis of the bias and bias correction of realized variance. Second, the heterogeneity of the Poisson process serves to capture dependence in trade durations at high frequency and heteroskedasticity in return volatility at low frequency and allows for a study into the properties of realized variance under alternative sampling schemes.

In discussing the distribution of returns, it is useful to distinguish between calendar time returns, i.e. $R(t|\tau)$, and transaction returns, i.e. $R_{k,m} = P_k - P_{k-m}$. It is straightforward to show that $R_{k,m}$ is Gaussian and $R_{k,1}$ has a first order autocovariance that is equal to $-\sigma_\nu^2$. While the serial correlation is desirable, Gaussianity is clearly not, at least for small values of m. In particular, it is well documented that single transaction returns (i.e. $m = 1$) are highly non-normal, in part due to price discreteness. However, when transaction returns are aggregated their distribution often converges rapidly to a normal (Ané and Geman 2000) suggesting that the model is reasonably well specified at all but the highest sampling frequencies. In contrast, the distribution of calendar time returns implied by the model is certainly not normal. Appendix A lists some moment expressions that can be derived for the restricted CPP-MA(1) model. The following notation is used:

$$\lambda_j = \int_{t_j-\tau_j}^{t_j} \lambda(u)\,du \qquad \text{and} \qquad \lambda_{i,j} = \int_{t_i}^{t_j-\tau_j} \lambda(u)\,du, \qquad (3)$$

---

[1]Although ε and ν are assumed to be independent, the analysis below can be extended straightfwardly to the case where they are correlated. Hansen and Lunde (2004a) emphasize the importance of allowing for such a correlation because it can lead to positive serial correlation of returns which may be needed for some types of data. The model considered here, however, can generate positive serial correlation of returns without this "leverage effect" through the higher order MA dependence on the noise component. Moreover, since the empirical analysis is concerned with transaction data (which are negatively correlated) and the correlation between ε and ν does not enrich the theoretical analysis, I ignore it here.

for $t_j \geq t_i + \tau_j$. Inspection of the moments suggests that the returns over fixed time intervals can have a fat tailed marginal distribution and are serially correlated. To provide some further insights here, consider the case where the intensity process is constant and integrates to λ_τ over intervals of length τ. The autocovariance function of returns is then given by

$$E\left[R\left(t|\tau\right)R\left(t+k\tau|\tau\right)\right] = -e^{-(k-1)\lambda_\tau}\left(1-e^{-\lambda_\tau}\right)^2\sigma_v^2,$$

which corresponds to the autocovariance function of an ARMA(1,1) process with AR parameter equal to $e^{-\lambda_\tau}$. Serial correlation of returns over short time intervals can thus be substantial, even at orders higher than one.

2.2 Definition of Alternative Sampling Schemes

An important contribution this paper makes is the analysis of realized variance under alternative sampling schemes. As an illustration, consider a trading day on which 7,800 transactions have occurred between 9.30am and 4.00pm. When the price process is sampled in calendar time at regular intervals of say 5 minutes, this will yield 78 return observations. An alternative, and as it will turn out more efficient, sampling scheme records the price process every time a multiple of 100 transactions have been executed. Although this "business time" sampling scheme leads to the same number of return observations, the properties of the resulting realized variance measure turn out to be fundamentally different. Crucially, I can show that in many cases sampling in business time achieves the minimal MSE of realized variance among all conceivable sampling schemes.

To formalize these ideas and facilitate discussion, I define three sampling schemes, namely general time sampling (GTS), calendar time sampling (CTS), and business time sampling (BTS). Both CTS and BTS are special cases of GTS.

Definition General Time Sampling Under GTS_N, the price process is sampled at time points $\left\{t_0^g, \ldots, t_i^g, \ldots, t_N^g\right\}$ over the interval $[t_0, t_0 + T]$ such that $t_0^g = t_0$, $t_N^g = t_0 + T$, and $t_i^g < t_{i+1}^g$.

Definition Calendar Time Sampling Under CTS_N, the price process is sampled at equidistantly spaced points in calendar time over the interval $[t_0, t_0 + T]$, i.e. $t_i^c = t_0 + i\delta$ for $i = \{0, \ldots, NT\}$ where $N = \delta^{-1}$ and δ denotes the sampling interval.

Definition Business Time Sampling Under BTS_N, the price process is sampled at equidistantly spaced points in business time over the interval $[t_0, t_0 + T]$, i.e. t_i^b for $i = \{0, \ldots, NT\}$ such that $t_0^b = t_0$, $t_N^b = t_0 + T$ and

$$\int_{t_i^b}^{t_{i+1}^b} \lambda(u)du = \frac{1}{NT}\int_{t_0}^{t_0+T}\lambda(u)du \equiv \lambda_N.$$

Before proceeding, a few remarks are in order. Notice that when the intensity process is latent, the BTS scheme is infeasible due to the specification of the sampling points. In the empirical analysis I therefore adopt a "feasible" BTS scheme which samples the price process every time a multiple of $\hat{\lambda}_N$ transactions have occurred, where $\hat{\lambda}_N$

FIGURE 1: BIAS OF RVAC UNDER BTS AND CTS
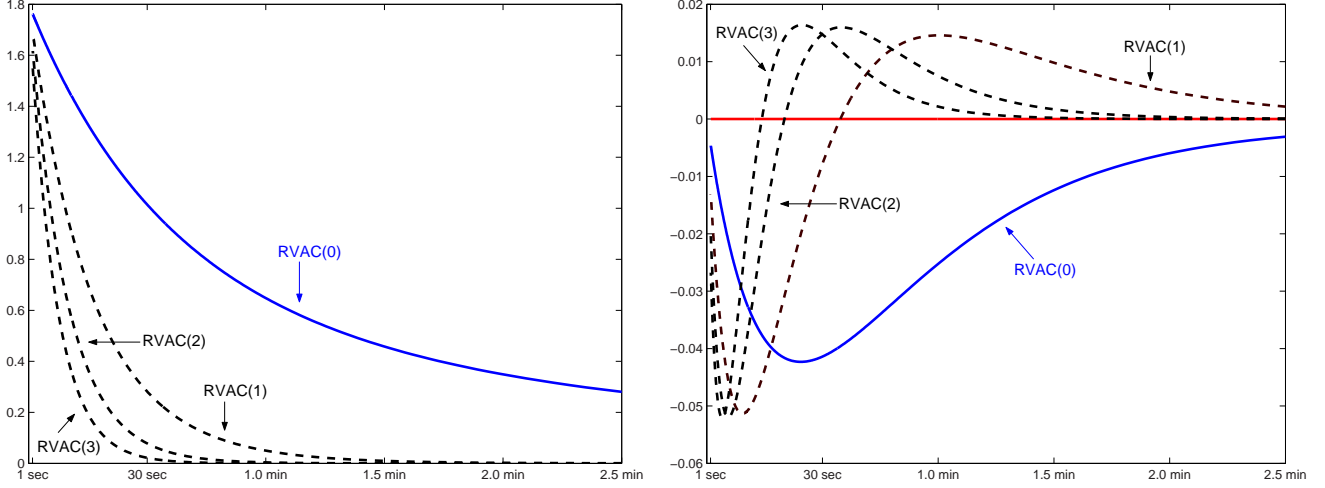
Note. The relative bias of RVAC under BTS for different values of q (left panel) and the difference in relative bias of RVAC under CTS and BTS (right panel). The restricted CPP-MA(1) parameters are set as $\lambda_{(0,1)} = 1000$, $\sigma_\varepsilon = \sqrt{5e-8}$, $\sigma_\nu = \sqrt{9e-8}/\sqrt{2}$, and the intensity process is specified as $\lambda(t) = (1 + \cos(2t\pi)/2)\lambda_{(0,1)}$.

is an unbiased estimator of λ_N given by the total number of transactions divided by NT. Further, the sampling frequency under CTS is defined in terms of seconds or minutes while under BTS it is defined in terms of the number of transactions. To facilitate discussion, I adopt the CTS terminology for both sampling schemes. A sampling frequency of 5 minutes under BTS on a trading day from 9.30 until 16.00 (390 minutes) therefore corresponds to sampling the price process each time a multiple of h transactions have occurred, where h is equal to the total number of transactions divided by 78. Also, throughout the paper I assume that market opening hours are from 9.30 until 16.00 so that each day has 390 minutes of trading. Finally, I introduce the following notation for the integrated intensity process over the interval $[a, b]$: $\lambda_{(a,b)} = \int_a^b \lambda(u)\, du$ so that $\lambda_{(t_j-\tau_j, t_j)} = \lambda_j$.

2.3 Properties of Bias Corrected Realized Variance

2.3.1 The Bias.

In the absence of market microstructure noise, realized variance yields an unbiased estimate of the "integrated return variance", which for the compound Poisson process over the unit time interval is equal to $\lambda_{(0,1)}\sigma_\varepsilon^2$. This defines the quantity of interest. Conditional on the intensity process, the bias of RVAC under GTS can be expressed as:

$$E_\lambda\left[RVAC_{N,q}^g - \lambda_{(0,1)}\sigma_\varepsilon^2\right] = \sigma_\nu^2 \sum_{i=1}^N (1 - e^{-\lambda_i})(2 - \sum_{k=1}^q (e^{-\lambda_{i-k,i}}(1 - e^{-\lambda_{i-k}}) + e^{-\lambda_{i,i+k}}(1 - e^{-\lambda_{i+k}}))).$$

Keep in mind that λ_i depends on the sampling points t_i^g and t_{i-1}^g specified by the sampling scheme. Under BTS this bias expression simplifies to:

$$E_\lambda \left[RVAC_{N,q}^b - \lambda_{(0,1)}\sigma_\varepsilon^2 \right] = 2N\sigma_\nu^2 \left(1 - e^{-\lambda_N} \right) e^{-q\lambda_N}.$$

This expression is quite intuitive and highlights some important features of the bias corrected realized variance. First of all, an increase in the order of the bias correction leads to an exponential decrease in bias. This is illustrated graphically in the left panel of Figure 1. Further, when the sampling frequency is low, i.e. N relatively small and λ_N relatively large, the bias of realized variance is negligible regardless of the order of the bias correction. This is not surprising since the serial correlation of returns at low frequency is small (recall that under BTS returns have a autocovariance function of an ARMA(1,1) process with AR parameter equal to $e^{-\lambda_N}$). However, when N increases, the bias increases and the following approximation

$$E_\lambda \left[RVAC_{N,q}^b - \lambda_{(0,1)}\sigma_\varepsilon^2 \right] \overset{N\, \text{large}}{\approx} 2\sigma_\nu^2\lambda_{(0,1)} \left(1 - qN^{-1}\lambda_{(0,1)} \right)$$

illustrates that for the bias to remain constant, the order of the bias correction would need to grow proportional to the sampling frequency. If not, the relative bias (i.e. the bias divided by $\lambda_{(0,1)}\sigma_\varepsilon^2$) will tend to $2 \left(\sigma_\nu/\sigma_\varepsilon \right)^2$ when the sampling frequency tends to infinity. This can also be observed from the bias curve in Figure 1 which asymptotes to its theoretical value of 1.8.

Another point of interest is to compare the bias of RVAC under alternative sampling schemes. When $q = 0$, it can be shown that the bias of realized variance under BTS is strictly larger than the bias of realized variance under CTS irrespective of the intensity process (Oomen 2004). When $q > 0$, it is not clear anymore which sampling scheme dominates and I therefore parameterize the intensity process as a simple deterministic function of time and compute the difference in bias under CTS and BTS to provide some illustration. The results are reported in the right panel of Figure 1. It is clear that BTS leads to a smaller bias of RVAC at relatively low frequencies, while CTS leads to a smaller bias of RVAC at relatively high frequencies. The difference in bias among the two sampling schemes is, however, very small in relation to the (relative) bias plotted in the left panel. For example, the bias under BTS is about 100% at a 10 second frequency for the first order bias corrected realized variance while the difference in bias among BTS and CTS is only about 5%. Although these specific values clearly depend on the parametrization of the model, qualitatively they appear robust to a wide range of parameter values and specifications of the intensity process.

Finally, it is noted that in the extreme case where either $N = 1$ or $N \rightarrow \infty$, all sampling schemes coincide. When $N = 1$ a single return is sampled over the unit time interval and it does therefore not depend on the way that this is done. When $N \rightarrow \infty$ both the timing and magnitude of *all* price innovations will be identified, irrespective of the sampling scheme, since the compound Poisson process is of finite variation. The bias, and also the MSE, are therefore the same under BTS, CTS, and GTS in these cases.

2.3.2 The Mean Squared Error.

In the absence of market microstructure noise the MSE of realized variance is equal to its variance, which can be expressed under GTS as follows:

$$MSE\left[RVAC_{N,0}^g\right] = \sum_{i=1}^{N}\left(3\sigma_\varepsilon^4\lambda_i^2 + 3\sigma_\varepsilon^4\lambda_i\right) + 2\sigma_\varepsilon^4\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\lambda_i\lambda_j - \lambda_{(0,1)}^2\sigma_\varepsilon^4.$$

Under BTS the MSE simplifies to:

$$MSE\left[RVAC_{N,0}^b\right] = \left(2\lambda_{(0,1)}\sigma_\varepsilon^2/N + 3\sigma_\varepsilon^2\right)\lambda_{(0,1)}\sigma_\varepsilon^2.$$

From this is can be seen that the realized variance is an inconsistent estimator of the integrated variance of the process. In particular, when $N \to \infty$, the variance of the estimator tends to $3\lambda_{(0,1)}\sigma_\varepsilon^4$. This highlights the fundamental difference between a pure jump process of finite variation and a diffusion based model for which the realized variance is generally consistent. Comparing MSE's under both sampling schemes reveals another interesting point:

$$MSE\left[RVAC_{N,0}^g\right] - MSE\left[RVAC_{N,0}^b\right] = 2\sigma_\varepsilon^4\sum_{i=1}^{N}\vartheta_i^2 > 0 \quad \text{where} \quad \vartheta_i = \int_{t_i-\tau_i}^{t_i}\lambda\left(u\right)du - \lambda_N.$$

When the intensity process varies over time, either stochastically or deterministically, BTS leads to a lower MSE of realized variance than any other conceivable sampling scheme. So despite the fact that the bias is maximized under BTS, it appears that this is more than offset by a reduction in variance of the estimator leading to a strictly lower MSE. Moreover, the benefits associated with BTS increase with an increase in the variability of the intensity process. See Oomen (2004) for further discussion and some empirical results.

Unfortunately, none of the above results can be established such generality when market microstructure noise is present. Depending on the parameter values and sampling frequency, BTS may either lead to a lower or a higher MSE of (bias corrected) realized variance than CTS. Numerical analysis below, however, shows that much of the above intuition for the simple case remains for all but a few extreme cases. In order to illustrate this, I compute the MSE of realized variance and first order bias corrected realized variance under BTS and CTS for the restricted CPP-MA(1) model and a wide range of parameter values. Under BTS, the MSE expression of the bias corrected realized variance can be obtained a function of the model parameters, λ_N, N, and q and can therefore be evaluated quickly. Because the expression is quite involved and lengthy, it is omitted to conserve space. The MSE calculation of bias corrected realized variance under CTS, or GTS for that matter, does not simplify and needs to be evaluated in a loop of $\left(\left(q+1\right)N\right)^2$ iterations. To keep computing time at reasonable levels, I perform all the calculations under CTS up to a sampling frequency of 1 minute and bias correction of up to order $q = 1$.

Figure 2 reports the percentage increase in MSE of realized variance (left panel) and bias corrected realized variance (right panel) under CTS relative to BTS. I refer to this quantity as "CTS loss" throughout the remainder

FIGURE 2: INEFFICIENCY OF CTS FOR BIAS CORRECTED REALIZED VARIANCE
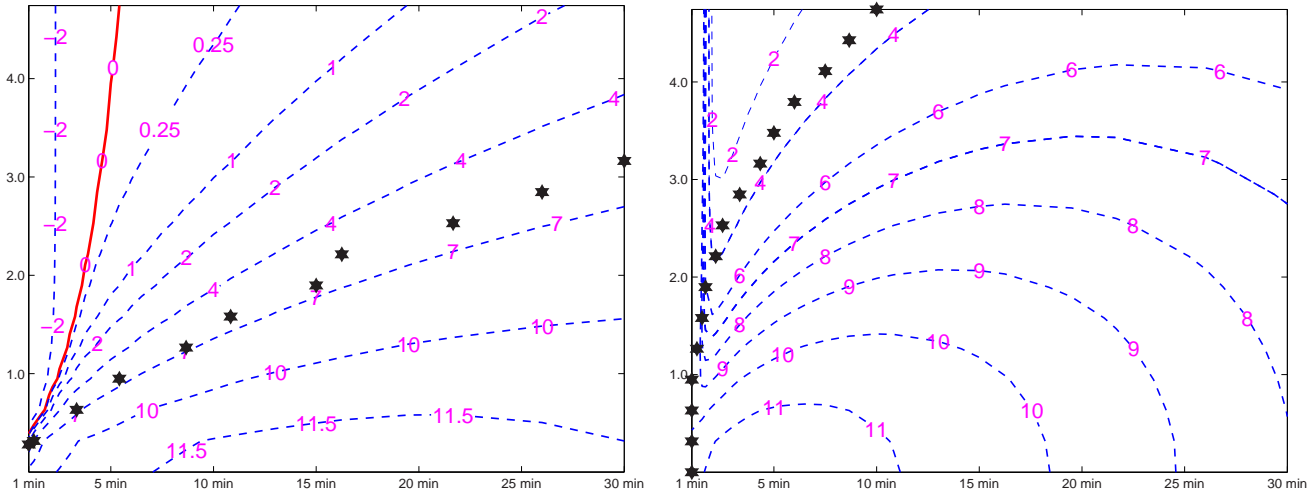
Note. Percentage CTS loss across sampling frequency (horizonal axis) and noise ratio (vertical axis) for realized variance (left panel) and first order bias corrected realized variance (right panel). The asterisks represent the sampling frequency which minimizes the MSE for a given level of market microstructure noise. The restricted CPP-MA(1) parameters are given in Figure 1.

of the paper. On the horizonal axis is the sampling frequency, and on the vertical axis is the so-called noise ratio, i.e. $\sigma_\nu/\sigma_\varepsilon$, which measures the magnitude of market microstructure noise relative to the "efficient" price innovation variance. The intensity process, σ_ε, and $\lambda_{(0,1)}$ are as specified above (see Figure 1). For realized variance, BTS outperforms CTS in terms of MSE for all combinations of sampling frequency and noise ratio that lie to the right of the solid curve in the left panel of Figure 2. The reduction in MSE under BTS can be as large as 11% for low levels of noise and sampling frequency, while CTS may lead to a lower MSE only when the noise ratio and the sampling frequency are extremely high. Hence, in practice, the performance of alternative sampling schemes crucially relies on the frequency at which the data is sampled. To address this issue, Figure 2 indicates the sampling frequency that minimizes the MSE of realized variance for a given level of noise ratio, i.e. the so called optimal sampling frequency, by an asterisks. In many cases, this will be the frequency of primary interest. Interestingly, the results indicate that BTS always outperforms CTS in the parameter region around the optimal sampling frequency, suggesting that for many practically relevant frequencies BTS is the superior sampling scheme. The empirical analysis below, confirms this.

The right panel of Figure 2 reports the corresponding results for the first order bias corrected realized variance. The first point to note is that when the bias correction is applied, the optimal sampling frequency increases dramatically. For example, for a noise ratio of 2.5, the optimal sampling frequency without bias correction is about 20 minutes while it increases to about 2 minutes with a first order bias correction. These results are in line with Hansen and Lunde (2004b). Further, it is clear that with the bias correction BTS outperforms CTS over an even larger region of sampling frequency and noise ratio combinations. More importantly, BTS remains the superior sampling scheme around the optimal sampling frequency range that is relevant for most practical applications.

9

FIGURE 3: LOG MSE OF BIAS CORRECTED REALIZED VARIANCE
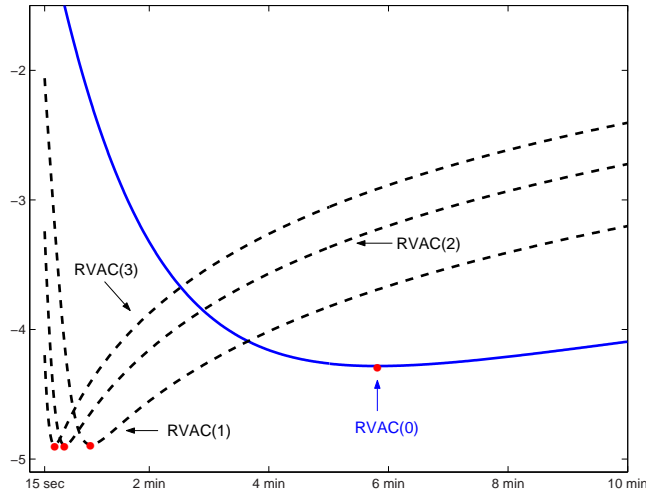
Note. logarithmic MSE of the bias corrected realized variance under BTS against the sampling frequency for different values of q. The price process is governed by the restricted CPP-MA(1) model with parameter values as given in Figure 1.

To conclude this section, I study the higher order bias corrections, i.e. $q > 1$, for the restricted CPP-MA(1) model with parameter values as before (see Figure 1). Figure 3 plots the logarithmic MSE of the bias corrected realized variance under BTS against the sampling frequency for different values of q. A couple of interesting findings emerge. Firstly, for a fixed sampling frequency, an increase in the order of the bias correction leads to a deterioration of MSE: the reduction in bias is more than offset by an increase in the variance of the estimator due to the incorporation of the additional correction term. Secondly, the introduction of a first order bias correction leads to a dramatic increase in optimal sampling frequency together with an associated decrease in MSE. In particular, without a bias correction, the optimal sampling frequency lies just under 6 minutes (i.e. 349 seconds) and the MSE of realized variance is 1.38%. With a first order bias correction, the optimal sampling frequency increases to 61 seconds at which the MSE is reduced to 0.75%, a decrease in MSE of more than 40%. Finally, and perhaps most striking, it appears that further increases in the order of the bias correction leads to a further increase in optimal sampling frequency but leaves the minimum MSE virtually unchanged! For $q = 2$ and $q = 3$, the optimal sampling frequency is 35 and 25 seconds respectively with a minimum MSE of 0.75% for both cases.

Even though the numerical analysis presented above inevitably required a specific choice of parameter values and characterization of the intensity function, extensive experimentation suggests that the results and main intuition remain unchanged for alternative sets of parameters. The analysis of the bias corrected realized variance under different model specification, possibly allowing for higher order dependence in the noise component, fat tailed t-innovations, and a leverage effects between ε and ν, is left for future research.

3 Optimal Sampling and Bias Correction in Practice

In this section I will discuss how the above methodology can be used in practice to (i) determine the optimal sampling frequency of a price process on a daily basis, (ii) measure the benefits of sampling in business time, and (iii) measure the benefits of bias correcting the realized variance. The analysis will be based on the restricted CPP-MA(1) model using both IBM and S&P500 Spider (SPY) transaction data. Overall the empirical findings confirm the above analysis. In particular, I find that a first order bias correction leads to a substantial reduction in MSE. In addition, sampling returns in business time as opposed to the common practice of sampling in calendar time further reduces the MSE on all days considered. For IBM (SPY) the average daily optimal sampling frequency without a bias correction lies around 2.5 (2.0) minutes while with a bias correction this increases to 13 (6) seconds. The average decrease in MSE associated with the first order bias correction exceeds 65% for both securities! The reduction in MSE due to business time sampling is much lower at around 3%-5%.

The Data IBM and SPY transaction data is retrieved from the TAQ database over the period from January 2, 2003 until August 31, 2003 (166 days). Data from all exchanges are considered but we discard transactions which take place before 9.45 and after 16.00. I also filter the data for instantaneous price reversals using an algorithm described in Oomen (2004) which removes 320 (1295) "outliers" for IBM (SPY). The cleaned data set for IBM (SPY) contains 1,224,127 (4,048,665) transactions.

Parameter Estimation For the restricted CPP-MA(1) model, the following parameters need to be estimated $\{\sigma_\varepsilon, \sigma_\nu, \lambda_N, \lambda(t)\}$. Using the first moment and first order autocovariance expressions of transaction returns I estimate the variance of the noise component and efficient price innovations, i.e.

$$
\begin{aligned}
Cov\left(R_k, R_{k-1}\right) &= -\sigma_\nu^2 \\
Var\left(R_k\right) &= \sigma_\varepsilon^2 + 2\sigma_\nu^2
\end{aligned}
$$

Although the autocovariance of transaction returns can in principle be positive, this does not occur on any given day in the data set. Nor, does $2\hat{\sigma}_\nu^2$ ever exceed the sample variance of transaction returns. The scaled integrated trade intensity, λ_N, can be estimated unbiasedly by the number of transactions on a given day divided by N. To obtain daily estimates of $\lambda(t)$, required to gauge the relative efficiency of CTS, I use a non-parametric smoothing technique described in Cowling, Hall, and Phillips (1996) and Diggle and Marron (1988). Bootstrap simulation results reported in Oomen (2004) indicate that this method provides very accurate estimates of the intensity process for realistic sample sizes.

Empirical Results Table 1 reports the average daily model parameter estimates (second column) plus the optimal sampling frequency, bias, and MSE for realized variance (third column), and first order bias corrected realized variance (fourth column) under BTS. Also the CTS loss is calculated for both estimators but to reduce computing

11

Table 1: Bias Corrected Realized Variance for IBM and S&P 500 SPIDER

	CPP-MA(1)			RVAC(0)				RVAC(1)			
	$\sigma_\nu/\sigma_\varepsilon$	σ_ε	$\lambda_{(0,1)}$	Freq.	Bias	MSE	CTSloss	Freq.	Bias	MSE	CTSloss
IBM											
Jan 03	1.66	1.27	8583	170	8.61	5.14	2.51	12	1.23	1.59	2.63
Feb 03	1.51	1.38	7698	162	8.35	5.57	1.99	13	1.41	1.76	2.71
Mar 03	1.46	1.42	8408	146	7.95	7.41	2.16	11	1.42	2.27	2.85
Apr 03	1.48	1.25	7772	156	8.22	3.43	3.09	12	1.43	1.09	4.14
May 03	1.35	1.14	7391	143	7.90	2.09	2.30	12	1.59	0.67	3.28
Jun 03	1.37	1.25	7053	150	8.05	3.08	3.24	13	1.60	1.02	4.15
Jul 03	1.22	1.34	6203	140	7.78	2.67	2.19	14	1.79	0.96	3.55
Aug 03	1.05	1.15	5907	119	7.20	1.25	2.66	14	1.93	0.44	4.71
Jan 03 - Aug 03	**1.39**	**1.28**	**7377**	**148**	**8.01**	**3.83**	**2.52**	**13**	**1.55**	**1.22**	**3.50**
SPY											
Jan 03	2.24	0.74	19666	147	8.01	2.94	5.72	7	0.41	0.76	3.64
Feb 03	2.23	0.78	23454	130	7.51	4.47	5.64	6	0.38	1.08	3.67
Mar 03	2.15	0.83	27747	112	6.98	9.13	5.24	5	0.41	2.00	3.56
Apr 03	2.13	0.70	24087	120	7.24	2.51	6.95	6	0.47	0.61	4.12
May 03	2.18	0.60	22819	128	7.48	1.37	6.56	6	0.43	0.34	4.37
Jun 03	1.99	0.59	24467	108	6.87	1.21	6.59	5	0.57	0.29	5.20
Jul 03	1.72	0.75	28043	82	5.95	8.29	4.52	4	0.69	2.76	3.76
Aug 03	1.78	0.55	24736	95	6.42	1.03	4.77	5	0.72	0.23	3.99
Jan 03 - Aug 03	**2.05**	**0.69**	**24389**	**115**	**7.05**	**3.86**	**5.75**	**6**	**0.51**	**1.01**	**4.04**

Note. The table reports the average daily estimates of the noise ratio ($\sigma_\nu/\sigma_\varepsilon$), efficient price innovation volatility ($\sigma_\varepsilon \times 10,000$), optimal sampling frequency ("Freq" in seconds), relative bias ("Bias" $\times 100$), mean squared error ("MSE" $\times 100$), and percentage CTS loss ("CTSloss") for the realized variance ("RVAC(0)") and first order bias corrected realized variance ("RVAC(1)") based on IBM and S&P500 Spider (SPY) transaction data.

time this figure is only computed once every month based on the average model parameters and intensity function estimates (i.e. for SPY over July 2003, to compute the MSE under CTS we need $(2 \times 23400/4)^2 > 136$ million iterations!).

The estimation results reveal a steady decrease in the noise component contribution to the price innovation over time. This is accompanied by a corresponding increase in optimal sampling frequency. Even though the average estimates for σ_ε differ significantly for IBM and SPY they imply a similar annualized return volatility due to the much higher volume of trading for SPY. Over the first 8 months of 2003, the model estimates a return

volatility of 17.38% for IBM and 17.04% for SPY which appears reasonable. Note that these estimates are model-based and use *all* transactions available.

For realized variance without the bias correction, the optimal sampling frequency of returns for IBM is about 2.5 minutes and for SPY it is about 2 minutes. These number are roughly in line with results reported in Bandi and Russell (2003) for IBM mid-quote data over a different period. As mentioned above, there is a clear upward trend in optimal sampling frequency due to a steady reduction in market microstructure noise. The average bias of realized variance at the optimal sampling frequency for both securities is around 7%-8% and the MSE around 4% (this is the MSE of percentage returns, i.e. $100 \times R$). The inefficiency of CTS, as measured by CTS loss, is about 2.5% for IBM and more than double for SPY.

The introduction of a first order bias correction leads, as theory suggests, to a substantial increase in optimal sampling frequency: from 2.5 minutes to 12 second for IBM and from 2.0 minutes to 6 seconds for SPY. More importantly, both the bias and the MSE of realized variance are dramatically reduced with an average fall of MSE of over 65%! These results are directly in line with Hansen and Lunde (2004a) who also find that the first order bias correction works extremely effectively. In comparison to these number, the benefits associated with sampling in business time are somewhat swamped. Still, BTS does reduce the MSE relative to CTS for each month considered here which underlines its superiority when returns are sampled at or around the optimal sampling frequency.

4   Concluding Remarks

I have studied some of the statistical properties of a bias corrected realized variance measure. Both the empirical and theoretical results are in agreement with closely related work by Hansen and Lunde (2004a). The paper is distinguished from existing literature in a number of ways. First, the analysis presented here is based on a pure jump process for high frequency security prices. As such, the results complement existing ones and widen the focuss of processes considered. Second, the impact that a particular choice of sampling scheme has on the properties of (bias corrected) realized variance is studied in some details. To the best of my knowledge this is the first paper to consider such issues. Finally, the paper provides a simple framework which can be used to gauge the statistical properties of realized variance and determine the optimal sampling frequency on a day-to-day basis. The two main results which emerge from the empirical and theoretical analysis is that (i) business time sampling is generally superior to the common practice of calendar time sampling and (ii) a first order bias correction leads to a dramatic improvement of the statistical properties of realized variance since it allows for efficient use of much higher frequency data.

Finally, it is noted that so far only the simplest CPP-MA(1) model has been considered. Although this model appears sufficiently flexible for the purpose at hand, the analysis of mid-quote data is likely to require a correlation between the market microstructure noise and the efficient price process or a higher order MA dependence structure in order to capture positive serial correlation of returns. This issue is left for future research.

A   Moments of Restricted CPP-MA(1)

Based on the characteristic function (omitted to conserve space but available from the author) for the price process given in expression (1), with $\mu_\varepsilon = \mu_v = 0$ and $\rho_0 = -\rho_1 = 1$, the following moment expressions can be derived:

$$
\begin{aligned}
\mathcal{M}_{2000}(\lambda_i) &= \lambda_i \sigma_\varepsilon^2 + 2(1 - e^{-\lambda_i})\sigma_\nu^2 \\
\mathcal{M}_{1100}(\lambda_i, \lambda_j, \lambda_{ij}) &= -e^{-\lambda_{ij}}(1 - e^{-\lambda_i})(1 - e^{-\lambda_j})\sigma_\nu^2 \\
\mathcal{M}_{4000}(\lambda_i) &= 3\lambda_i(1 + \lambda_i)\sigma_\varepsilon^4 + 12\sigma_\nu^4(1 - e^{-\lambda_i}) + 12\sigma_\varepsilon^2 \lambda_i \sigma_\nu^2 \\
\mathcal{M}_{2200}(\lambda_i, \lambda_j, \lambda_{ij}) &= \lambda_i \lambda_j \sigma_\varepsilon^4 + 2(\lambda_i + \lambda_j - \lambda_i e^{-\lambda_j} - \lambda_j e^{-\lambda_i})\sigma_\nu^2 \sigma_\varepsilon^2 \\
&\quad + 2(1 - e^{-\lambda_j})(2 + e^{-\lambda_{ij}})(1 - e^{-\lambda_i})\sigma_\nu^4 \\
\mathcal{M}_{3100}(\lambda_i, \lambda_j, \lambda_{ij}) &= -3e^{-\lambda_{ij}}(1 - e^{-\lambda_j})(\lambda_i \sigma_\varepsilon^2 + 2(1 - e^{-\lambda_i})\sigma_\nu^2)\sigma_\nu^2 \\
\mathcal{M}_{1210}(\lambda_i, \lambda_j, \lambda_k, \lambda_{ij}, \lambda_{jk}) &= 2e^{-\lambda_{ij} - \lambda_{jk}}(1 - e^{-\lambda_i})(1 - e^{-\lambda_j})(1 - e^{-\lambda_k})\sigma_\nu^4 \\
\mathcal{M}_{2110}(\lambda_i, \lambda_j, \lambda_k, \lambda_{jk}) &= -e^{-\lambda_{jk}}(1 - e^{-\lambda_j})(1 - e^{-\lambda_k})(\lambda_i \sigma_\varepsilon^2 + 2(1 - e^{-\lambda_i})\sigma_\nu^2)\sigma_\nu^2 \\
\mathcal{M}_{1111}(\lambda_i, \lambda_j, \lambda_k, \lambda_m, \lambda_{ij}, \lambda_{km}) &= e^{-\lambda_{ij} - \lambda_{km}}(1 - e^{-\lambda_i})(1 - e^{-\lambda_j})(1 - e^{-\lambda_k})(1 - e^{-\lambda_m})\sigma_\nu^4
\end{aligned}
$$

where

$$
\mathcal{M}_{p_i p_j p_k p_m}(\lambda_i, \lambda_j, \lambda_k, \lambda_m, \lambda_{ij}, \lambda_{jk}, \lambda_{km}) = E\left[R(t_i|\tau_i)^{p_i} R(t_j|\tau_j)^{p_j} R(t_k|\tau_k)^{p_k} R(t_m|\tau_m)^{p_m}\right]
$$

for $i < j < k < m$

References

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens, 2001, "The Distribution of Stock Return Volatility," *Journal of Financial Economics*, 61, 43–76.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys, 2000, "Great Realizations," *Risk*, pp. 105–108.

——— , 2003, "Modeling and Forecasting Realized Volatility," *Econometrica*, 71 (2), 579–625.

Ané, T., and H. Geman, 2000, "Order Flow, Transaction Clock, and Normality of Asset Returns," *Journal of Finance*, 55(5), 2259–2284.

Bandi, F. M., and J. R. Russell, 2003, "Microstructure Noise, Realized Volatility, and Optimal Sampling," manuscript GSB, The University of Chicago.

Barndorff-Nielsen, O. E., and N. Shephard, 2004, "Econometric Analysis of Realised Covariation: High Frequency Based Covariance, Regression and Correlation in Financial Economics," *forthcoming Econometrica*.

Corsi, F., G. Zumbach, U. A. Müller, and M. Dacorogna, 2001, "Consistent High-Precision Volatility from High-Frequency Data," Olsen Group Working Paper.

Cowling, A., P. Hall, and M. J. Phillips, 1996, "Bootstrap Confidence Regions for the Intensity of a Poisson Point Process," *JASA*, 91 (436), 1516–1524.

Diggle, P., and J. Marron, 1988, "Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation," *JASA*, 83 (403), 793–800.

Easley, D., N. M. Kiefer, M. O'Hara, and J. B. Paperman, 1996, "Liquidity, Information, and Infrequently Traded Stocks," *Journal of Finance*, 51 (4), 1405–1436.

French, K. R., G. W. Schwert, and R. Stambaugh, 1987, "Expected Stock Returns and Volatility," *Journal of Financial Economics*, 19, 3–29.

Froot, K. A., and A. F. Perold, 1995, "New Trading Practices and Short-Run Market Efficiency," *Journal of Futures Markets*, 15(7), 731–765.

Hansen, P. R., and A. Lunde, 2004a, "An Unbiased Measure of Realized Variance," manuscript Department of Economics, Brown University.

———, 2004b, "Realized Variance and IID Market Microstructure Noise," manuscript Brown University.

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

Oomen, R. C., 2003, "Three Essays on the Econometric Analysis of High Frequency Financial Data," Ph.D. thesis, European University Institute.

———, 2004, "Properties of Realized Variance for a Pure Jump Process: Calendar Time Sampling versus Business Time Sampling," manuscript Warwick Business School.

Press, S. J., 1967, "A Compound Events Model for Security Prices," *Journal of Business*, 40(3), 317–335.

Protter, P., 1990, *Stochastic Integration and Differential Equations: A new approach*. Springer-Verlag, Berlin.

Roll, R., 1984, "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market," *Journal of Finance*, 39(4), 1127–1139.

Zhang, L., P. A. Mykland, and Y. Ait-Sahalia, 2003, "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data," manuscript Princeton University.

Zhou, B., 1996, "High Frequency Data and Volatility in Foreign-Exchange Rates," *Journal of Business & Economic Statistics*, 14(1), 45–52.