Properties of Realized Variance for a Pure Jump Process: Calendar Time Sampling versus Business Time Sampling

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

March 2004
Comments are welcome

Abstract

In this paper we study the impact of market microstructure effects on the properties of realized variance using a pure jump process for high frequency security prices. Closed form expressions for the bias and mean squared error of realized variance are derived under alternative sampling schemes. Importantly, we show that business time sampling is generally superior to the common practice of calendar time sampling in that it leads to a reduction in mean squared error. Using IBM transaction data we estimate the model parameters and determine the optimal sampling frequency for each day in the data set. The empirical results reveal a downward trend in optimal sampling frequency over the last 4 years with considerable day-to-day variation that is closely related to changes in market liquidity.

Keywords: Market Microstructure Noise; Bias and MSE of Realized Variance; Time Deformation; Optimal Sampling

*Roel Oomen is also a research affiliate of the Department of Quantitative Economics at the University of Amsterdam, The Netherlands. The initial draft of this paper has been prepared for the "Analysis of High-Frequency Financial Data and Market Microstructure" conference, December 15-16, 2003, Taipei. The author thanks Frank Diebold (discussant), Peter Hansen, Douglas Steigerwald, and the conference participants at the January 2004 North American meeting of the Econometric Society for helpful comments. The author also thanks Michael Boldin for providing the TAQ data.

1 Introduction

The recent trend towards the model-free measurement of asset return volatility has been spurred by an increase in the availability of high frequency data and the development of rigorous foundations for realized variance, defined as the sum of squared intra-period returns (see for example Andersen, Bollerslev, Diebold, and Labys (2003), Barndorff-Nielsen and Shephard (2004b), Meddahi (2002)). While the theory suggest that the integrated variance can be estimated arbitrarily accurate by summing up squared returns at sufficiently high frequency, the validity of this result crucially relies on the price process conforming to a semi-martingale thereby ruling out any type of non-trivial dynamics in the first moment of returns. Yet, it is well known that the various market microstructure effects induce serial correlation in returns that are sampled at the highest frequencies. This apparent conflict between the theory and practice of model-free volatility measurement is what motivates this paper. Here, we propose a framework which allows us to study the statistical properties of realized variance in the presence of market microstructure noise. In particular, we derive closed form expressions for the bias and mean squared error (MSE) of realized variance and establish their relation with the sampling frequency. The principal contribution this paper makes is that the analysis explicitly distinguishes among different sampling *schemes* which, to the best of our knowledge, has not yet been considered in the literature. Importantly, both the theoretical and the empirical results suggest that the MSE of realized variance can be reduced by sampling returns on a business time scale as opposed to the common practice of sampling in calendar time.

Inspired by the work of Press (1967, 1968) we use a compound Poisson process to model the asset price as the accumulation of a finite number of *jumps*, each of which can be interpreted as a transaction return with the Poisson process counting the number of transactions. To capture the impact of market microstructure noise, we allow for a flexible MA(q) dependence structure on the price increments. Further, we leave the Poisson intensity process unspecified but point out that in practice both stochastic and deterministic variation in the intensity process may be needed in order to capture duration dependence at high frequency, persistence in return volatility at low frequency, and seasonality of market activity. Despite these alterations, the model remains analytically tractable in that the characteristic function of the price process can be derived in closed form after conditioning on the integrated intensity process.

The move towards a semi-parametric pure jump process with discontinuous sample paths of finite variation marks a significant departure from the popular domain of diffusion based models. For example, in our framework realized variance is an inconsistent estimator of the (jump analogue of) integrated variance. Also, in the presence of market microstructure noise, the bias of realized variance does not tend to infinity when the sampling frequency approaches zero as is often the case for a diffusive price process (see for example Bandi and Russell (2003), Zhang, Mykland, and Ait-Sahalia (2003)). Yet, despite such seemingly fundamental differences, it is important not to overstate these because both the pure jump process and the diffusive process can give rise to a number of similar results and intuition regarding the statistical properties of realized variance. The main motivation for using the jump model here, is that it provides a convenient framework in which to analyze the statistical

properties of realized variance under different *sampling schemes* in the presence of market microstructure noise. Furthermore, the model has great intuitive appeal and is in line with a number of recent papers which emphasize the important role that pure jump processes play in finance from a time series modelling point of view (e.g. Carr, Geman, Madan, and Yor (2002), Maheu and McCurdy (2004), Rydberg and Shephard (2003)) and an option pricing perspective (e.g. Carr and Wu (2003), Geman, Madan, and Yor (2001)).

On the theoretical side, we make three contributions in this paper. First, we provide a flexible and analytically tractable framework in which to rigorously investigate the statistical properties of realized variance in the presence of market microstructure noise. Closed form expressions for the bias and MSE of realized variance are derived, based on which the optimal sampling frequency can be determined. Second, the analysis explicitly distinguishes among different sampling *schemes*, including business time sampling and calendar time sampling. In many cases, a superior (and inferior) sampling scheme can be identified. As such, the results here provide new insights into the impact of a particular choice of sampling scheme on the properties of realized variance. Third, the modelling framework allows for a general MA(q) dependence structure on price increments which is important to capture non-iid market microstructure noise as emphasized by Hansen and Lunde (2004b). Further, on the empirical side, the paper provides an illustration of how the model can be used to determine the optimal sampling frequency in practice and gauge the efficiency of a particular sampling scheme.

The principal finding which emerges from both our theoretical and empirical analysis, is that business time sampling is generally superior to the common practice of calendar time sampling in that it leads to a reduction in MSE of realized variance. Intuitively, business time sampling (i.e. prices are sampled on a business time scale defined by for example the cumulative number of transactions as opposed to a calendar time scale defined in seconds) gives rise to a return series that is effectively "devolatized" through the appropriate deformation of the calendar time scale. It is precisely this feature of the sampling scheme which leads to the improvement in efficiency of the variance estimator. Further, we show that the magnitude of this efficiency gain increases with an increase in the variability of trading intensity suggesting that the benefits of sampling in business time are most pronounced on days with irregular trading patterns. Only in exceptional circumstances, where either the sampling frequency is extremely high and far beyond its "optimal" level or market microstructure noise is unrealistically dominant, calendar time sampling can lead to a slight improvement in MSE relative to business time sampling. The empirical analysis confirms the above. Using IBM transaction data from January 2000 until August 2003 we estimate the model parameters and trade intensity process to determine the MSE loss associated with calendar time sampling. We find that for each day in the sample, business time sampling leads to a lower MSE of realized variance. The average MSE reduction is a modest 3% but can be as high as 30%-40% on days with dramatic swings in market activity.

In addition to the analysis of alternative sampling schemes, the proposed model can also be used to deliver a time series of *daily* estimated optimal sampling frequencies. For the IBM transaction data, we find that the average optimal sampling frequency is about 3 minutes over the sample period (5 minutes in 2000 and 2.5 minutes in 2003). Simulations indicate that the impact of measurement error in the model parameters does not

bias the results here. Further, we find considerable variation in the optimal sampling frequency from day to day that is closely related to changes in market liquidity.

That market microstructure noise impacts on the statistical properties of realized variance is certainly not an observation original to this research. Andersen, Bollerslev, Diebold, and Labys (2000) were one of the first to explicitly document some anecdotal evidence on the relation between sampling frequency, market microstructure noise, and the bias of the realized variance measure. Now, a growing body of literature emphasizes the crucial role that the sampling of the price process and the filtering of market microstructure noise plays. The papers that are most closely related to the current one are by Bandi and Russell (2003) and Hansen and Lunde (2004b) who build on the work by Andersen, Bollerslev, Diebold, and Labys (2003) and Barndorff-Nielsen and Shephard (2004b) in a similar way that this paper builds on Press (1967). Somewhat surprisingly, many of the results and intuition are qualitatively comparable and can thus be seen to complement each other since they are derived under different assumptions on the price process. What differentiates our analysis and our results from theirs is that we explicitly distinguish among different sampling schemes. Also, our approach to determining the optimal sampling frequency is fundamentally different from Bandi and Russell (2003) and much less sensitive to measurement error.

To conclude, we emphasize that the search for an "optimal" sampling frequency or scheme is only one of many possible takes on the issue. Alternative approaches range from the explicit modelling of market microstructure noise in a fully parametric fashion proposed by Corsi, Zumbach, Müller, and Dacorogna (2001) to a non parametric Newey-West type adjustment as advocated by Hansen and Lunde (2004a). In a recent paper, Zhang, Mykland, and Ait-Sahalia (2003) discuss how subsampling can be used to reduce the impact of market microstructure noise while avoiding the need to aggregate or "throw away" data. Although this approach may have several advantages, we point out that the efficiency gains that can be achieved in practice crucially rely on the design of the grid over which the subsampling takes place. So even if subsampling is the preferred method for the application at hand, the results on the properties of realized variance under different sampling frequencies *and* schemes presented in this paper may well provide some useful guidance as to the design of an "optimal" grid structure.

The remainder of the paper is structured as follows. In Section 2, we introduce the extended compound Poisson process as a model for high frequency security prices. We discuss how the model can account for market microstructure noise, and derive the joint characteristic function of returns. Section 3 contains a theoretical discussion of the properties of realized variance under alternative sampling schemes in terms of bias and MSE. Section 4 outlines the estimation of the model and present the empirical results for the IBM transaction data. Section 5 concludes.

2 A Pure Jump Process for High Frequency Security Prices

Let the logarithmic asset price at time t, $P(t)$, follow a heterogeneous compound Poisson process with serially correlated increments, i.e.

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

where $\varepsilon_j \sim$iid $\mathcal{N}\left(\mu_\varepsilon, \sigma_\varepsilon^2\right)$, $\nu_j \sim$ iid $\mathcal{N}\left(\mu_\nu, \sigma_\nu^2\right)$, and $M(t)$ is a Poisson process with instantaneous intensity $\lambda(t) > 0$. Throughout the remainder of this paper we refer to the model in expression (1) as CPP-MA(q) when $\rho_q \neq 0$ and $\rho_k = 0$ for $k > q$ and CPP-MA(0) when $\rho_k = 0$ for $k \geq 0$. Since the focuss will be on the analysis of financial transaction data, we 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, our model is closely related to the literature on subordinated processed initiated by Clark (1973). Despite a long tradition in the statistics literature (see Andersen, Borgan, Gill, and Keiding (1993), Karlin and Taylor (1981) and references therein), pure jump processes have until recently received only moderate attention in finance, after being introduced by Press (1967, 1968) . Yet, a number of recent papers provide compelling evidence that these type of processes can be successfully applied to many issues in finance ranging from the modelling of low frequency returns (e.g. Maheu and McCurdy (2003, 2004)) and high frequency transaction data (e.g. Bowsher (2002), Rogers and Zane (1998) , Rydberg and Shephard (2003)), to the pricing of derivatives (e.g. Barndorff-Nielsen and Shephard (2004a), Carr and Wu (2003), Cox and Ross (1976), Geman, Madan, and Yor (2001), Mürmann (2001)). Also for our purpose here, it turns out that a pure jump process such as the CPP-MA(q) defined above is ideally suited to study the impact of market microstructure noise on the properties of realized variance.

It is well documented that market microstructure effects, such as non-synchronous trading, bid/ask bounce, screen fighting, etc, introduce serial correlation in returns at high frequency. It is precisely this what motivates the MA(q) structure on η from a statistical viewpoint. Yet, a closer look at the model in "business time" (i.e. clock runs on $M(t)$ as opposed to t for "calendar time") provides a more intuitive motivation. Let P_k denote the logarithmic price after the k^{th} transaction, i.e.

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

so that

$$R_{k,m} \equiv P_k - P_{k-m} = R_{k,m}^e + \sum_{j=0}^{m-1} \eta_{k-j}.$$

This suggests the following interpretation: the observed but "market microstructure noise contaminated" price, P, is equal to the efficient but unobserved price, P^e, plus an accumulated noise component. A similar interpretation

holds for returns. Note that the model, as it stands, has the undesirable feature that the contribution of the noise component towards the variance of returns does not necessarily diminish under temporal aggregation.

We therefore impose the following restriction on the MA(q) parameters $\eta_k = \sum_{i=0}^{q-1} \overline{\rho}_i \left(\nu_{k-i} - \nu_{k-i-1}\right)$ so that the variance of $R_{k,m} - R^e_{k,m} = \sum_{i=0}^{q-1} \overline{\rho}_i \left(\nu_{k-i} - \nu_{k-m-i}\right)$ only depends on q and not on m. We emphasize that a number of different restrictions can be imposed to achieve the same result. However, since the above restriction leads to first order negative serial correlation of returns which is needed later on in the empirical analysis, we adopt this particular specification of the model throughout the remainder of this paper and refer to it as the "restricted CPP-MA(q)".

In anticipation of the main theorem characterizing the distribution of returns, we introduce some further notation. Let $R\left(t_j|\tau_j\right) \equiv P\left(t_j\right) - P\left(t_j - \tau_j\right)$ denote the continuously compounded return over time interval $[t_j - \tau_j, t_j]$ with corresponding integrated intensity process λ_j measured over the same interval and defined as:

$$\lambda_j = \int_{t_j-\tau_j}^{t_j} \lambda\left(u\right) du \qquad \text{and} \qquad \lambda_{i,j} = \int_{t_i}^{t_j-\tau_j} \lambda\left(u\right) du \tag{2}$$

for $t_j \geq t_i + \tau_j$. Similarly, $\lambda_{i,j}$ measures the integrated intensity between the sampling intervals of $R\left(t_i|\tau_i\right)$ and $R\left(t_j|\tau_j\right)$. Keep in mind that the notation for the integrated intensity process is always associated with a sampling interval, i.e. λ_j is associated with both t_j and τ_j.

Theorem 2.1 *For the CPP-MA(q) price process given in expression (1), the joint characteristic function of returns conditional on the intensity process λ, i.e. $\phi_{RR}\left(\xi_1, \xi_2 | t_1, t_2, \tau_1, \tau_2, q\right) \equiv E_\lambda\left(\exp\left\{i\xi_1 R\left(t_1|\tau_1\right) + i\xi_2 R\left(t_2|\tau_2\right)\right\}\right)$ for $t_2 \geq t_1 + \tau_2$, is given by:*

$$\begin{aligned}
&e^{-\lambda_1-\lambda_2}\left(1 + \exp\left\{\xi_1^2 c_0 + \lambda_1 e^{c_1}\right\}\left(1 - \frac{\Gamma\left(q, \lambda_1 e^{c_1}\right)}{\Gamma\left(q\right)}\right) + \exp\left\{\xi_2^2 c_0 + \lambda_2 e^{c_2}\right\}\left(1 - \frac{\Gamma\left(q, \lambda_2 e^{c_2}\right)}{\Gamma\left(q\right)}\right)\right) \\
&+\Psi_q\left(\xi_1, \xi_2, q, \Omega_q\right) \exp\left\{\left(\xi_1^2 + \xi_2^2\right) c_0 + \lambda_1\left(e^{c_1} - 1\right) + \lambda_2\left(e^{c_2} - 1\right)\right\} \\
&+e^{-\lambda_1-\lambda_2}\left(\sum_{h=1}^{q-1} \exp\left\{i\xi_1 h\left(\mu_\varepsilon + \mu_v\overline{\rho}\right) - \frac{1}{2}\xi_1^2 \Sigma_q\left(h\right)\right\}\frac{\lambda_1^h}{h!} + \sum_{k=1}^{q-1} \exp\left\{i\xi_2 k\left(\mu_\varepsilon + \mu_v\overline{\rho}\right) - \frac{1}{2}\xi_2^2 \Sigma_q\left(k\right)\right\}\frac{\lambda_2^k}{k!}\right) \\
&+\sum_{m=0}^{\infty}\sum_{h=1}^{q-1}\sum_{k=1}^{\infty} \phi_q\left(h,k,m\right)\frac{\lambda_2^k}{k!e^{\lambda_2}}\frac{\lambda_1^h}{h!e^{\lambda_1}}\frac{\lambda_{1,2}^m}{m!e^{\lambda_{1,2}}} + \sum_{m=0}^{\infty}\sum_{h=q}^{\infty}\sum_{k=1}^{q-1} \phi_q\left(h,k,m\right)\frac{\lambda_2^k}{k!e^{\lambda_2}}\frac{\lambda_1^h}{h!e^{\lambda_1}}\frac{\lambda_{1,2}^m}{m!e^{\lambda_{1,2}}}
\end{aligned}$$

where Γ denotes the (incomplete) Gamma function,

$$\Psi_q\left(\xi_1, \xi_2, q, \Omega_q\right) = \left(1 - \frac{\Gamma\left(q, \lambda_1 e^{c_1}\right)}{\Gamma\left(q\right)}\right)\left(1 - \frac{\Gamma\left(q, \lambda_2 e^{c_2}\right)}{\Gamma\left(q\right)}\right)\left(1 - \frac{\Gamma\left(q, \lambda_{1,2}\right)}{\Gamma\left(q\right)} + \sum_{m=0}^{q-1} e^{-\xi_1\xi_2\Omega_q(m)}\frac{\lambda_{1,2}^m}{m!e^{\lambda_{1,2}}}\right)$$

and $\Sigma_q, \Omega_q,$, $\phi_q\left(h,k,m\right)$, and c_k are as defined by expressions (4), (5), (6), and (7), in Appendix A.

Proof See Appendix A.

The characteristic function given in Theorem 2.1 above allows for the straightforward derivation of conditional return moments and cross moments (or cumulants) of any order for all models within the class defined by expression (1). Conditioning on the (integrated) intensity process is done to avoid the need to specify a possibly

restrictive dynamic model for λ which would, in all probability, complicate the derivation of the characteristic function substantially. Moreover, since the interest in this paper lies in the realized variance calculations which are ex post in nature, and the trade intensity process can be estimated reasonably accurate in practice by non-parametric smoothing methods (discussed and implemented using IBM TAQ data in section 4), conditioning on intensity is justified.

Inspection of the pure jump model in expression (1), and the various return moments which can be derived for it, suggest that the price process in calendar time (i.e. t) is sufficiently flexible to capture a number of salient features of high frequency returns including, serial correlation, fat tailed marginal distributions, seasonal patterns in market activity, and conditional dependence in transaction durations at high frequency and return volatility at low frequency through the variation in the intensity process. For instance, the kurtosis of returns in calendar time for the CPP-MA(0) model is equal to $3 + 3\left(\int \lambda(u)du\right)^{-1}$ which can be quite substantial over short horizons when the corresponding integrated intensity is small. Further, a simple two factor model for the intensity process (with one quickly and one slowly mean reverting factor) is capable of generating dependence in both durations and return volatility. A less appealing feature of the model is that the distribution of returns in business time are Gaussian. This is clearly in strong disagreement with the data which appear to be more accurately described by a multinomial distribution at the transaction level due to price discreteness. On the positive side, we point out that the multinomial distribution converges rapidly to a normal under aggregation and that, as a consequence, the model misspecification in business time will diminish quickly and have no material impact on the results when we study the properties of the price process at lower (business time) frequencies. Since we concentrate on the range of frequencies which minimize the MSE, and the empirical results below indicate that these frequencies lie around 3 minutes in calendar time, or multiples of roughly 60 trades in business time on representative days, our approach is likely to suffer very little from model misspecification. The work by Ané and Geman (2000) further underlines the validity of our model in business time at all but the highest sampling frequencies.

3 Realized Variance: Business Time Sampling versus Calendar Time Sampling

In this section we investigate the statistical properties of realized variance based on the restricted CPP-MA(q) model. The analysis will be carried out along two dimensions, namely by varying the *sampling frequency* and by varying the *sampling scheme*. The impact of a change in sampling frequency on the statistical properties of realized variance is quite intuitive: an increase in sampling frequency will lead to (i) a reduction in the variance of the estimator due to the increase in the number of observations and (ii) an increase in the bias of the estimator due to the amplification of market microstructure noise induced return serial correlation. It is this tension between the bias and the variance of the estimator which motivates the search for an "optimal" sampling frequency, i.e. the frequency at which the MSE is minimized. The contribution that the paper makes on this front is the derivation of a closed form expression for the MSE of realized variance as a function of the sampling frequency which allows for the identification of the optimal sampling frequency. The results here are closely related to those of Bandi

and Russell (2003) and Hansen and Lunde (2004a) who obtain similar expressions within their framework. The second issue, i.e. the impact of a change in sampling *scheme* on the statistical properties of realized variance, is far less obvious and to the best of our knowledge this paper is the first to provide a comprehensive analysis. 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, we 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, we 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

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

A few remarks are in order. When the intensity process is latent, the BTS scheme is infeasible due to the specification of the sampling points. In the empirical analysis we therefore adopt a "feasible" BTS scheme which samples the price process every time a multiple of $\hat{\lambda}_N$ transactions have occurred, where $\hat{\lambda}_N$ is an unbiased estimator of λ_N given by the total number of transactions divided by NT. Further, without loss of generality, we will concentrate on the unit time interval (i.e. $t_0 = 0$ and $T = 1$) in the theoretical part and assume $\tau_i = t_i - t_{i-1}$ to ensure that returns are adjacent (without overlap or gaps, i.e. $R\left(t_i|\tau_i\right) = P\left(t_i\right) - P\left(t_{i-1}\right)$). We also introduce the following notation for the integrated intensity process over the interval $[a, b]$: $\lambda_{(a,b)} = \int_a^b \lambda\left(u\right) du$ so that $\lambda_{(t_j-\tau_j,t_j)} = \lambda_j$. The return variance of the efficient price process over the unit time interval can then be expressed as $\lambda_{(0,1)}\sigma_\varepsilon^2$, i.e. the (scaled) integrated intensity process. Throughout the remainder of this paper we will loosely refer to this quantity as the integrated variance process. Finally, we note that 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, we 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 (total number of trades / 78) transactions have occured.

3.1 Absence of Market Microstructure Noise

In the absence of market microstructure noise, realized variance is an unbiased estimator of the integrated variance, i.e. $E_\lambda\left[\sum_{i=1}^{N} R\left(t_i|\tau_i\right)^2\right] - \lambda_{(0,1)}\sigma_\varepsilon^2 = 0$, irrespective of the sampling scheme. The MSE is therefore given by the variance of the estimator which can be expressed under GTS as:

$$\begin{aligned} MSE\left(GTS_N\right) &= E_\lambda\left[\sum_{i=1}^{N} R\left(t_i|\tau_i\right)^4 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} R\left(t_i|\tau_i\right)^2 R\left(t_j|\tau_j\right)^2\right] - E_\lambda\left[\sum_{i=1}^{N} R\left(t_i|\tau_i\right)^2\right]^2 \\ &= \sum_{i=1}^{N}\left(3\sigma_\varepsilon^4\lambda_i^2 + 3\sigma_\varepsilon^4\lambda_i\right) + 2\sigma_\varepsilon^4\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\lambda_i\lambda_j - \lambda_{(0,1)}^2\sigma_\varepsilon^4 \end{aligned}$$

where the superscripts on t_i and τ_i are omitted for notational convenience. Also, keep in mind that λ_i is associated with t_i and τ_i. For the corresponding sampling scheme in calendar time, i.e. CTS_N, the MSE is given by the expression above where $\lambda_i = \int_{(i-1)\delta}^{i\delta}\lambda\left(u\right)du$ and $\delta = N^{-1}$. Finally, under BTS_N, the MSE expression simplifies to

$$MSE\left(BTS_N\right) = \left(2\lambda_{(0,1)}\sigma_\varepsilon^2/N + 3\sigma_\varepsilon^2\right)\lambda_{(0,1)}\sigma_\varepsilon^2$$

From this expression it is clear that the MSE does not tend to zero when $N \to \infty$, which reveals the inconsistency of the realized variance measure in our framework. Since the price path is assumed to follow a pure jump process of bounded variation, there will be a point beyond which increasing the sampling frequency does not generate any additional information. Instead, increasing the sampling frequency will just lead to more and more zeros being added into the sampled return series. Hence the inconsistency. This is clearly not the case for a diffusive price process which is of infinite variation and for which the realized variance measure is consistent.

Based on the above expressions for the MSE, we can now compare the relative efficiency of the realized variance measure under the alternative sampling schemes. In order to do so, we need to introduce some more notation:

$$\vartheta_i = \int_{t_i-\tau_i}^{t_i}\lambda\left(u\right)du - \lambda_N$$

The quantity ϑ_i measures the difference between the integrated intensity over the i^{th} sampling increment associated with GTS_N, i.e. $\left[t_{i-1}^g, t_i^g\right]$, and that of BTS_N, i.e. $\left[t_{i-1}^b, t_i^b\right]$. A graphical illustration of ϑ_i is given in the left panel of Figure 1. Notice that by definition $\sum_{i=1}^{N}\vartheta_i = 0$. The difference in MSE for GTS, relative to BTS, can now be expressed as:

$$\begin{aligned} MSE\left(GTS_N\right) - MSE\left(BTS_N\right) &= 3\sigma_\varepsilon^4\sum_{i=1}^{N}\left(\lambda_i^2 - \lambda_N^2\right) + 2\sigma_\varepsilon^4\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\lambda_i\lambda_j - \lambda_N^2\right) \\ &= 2\sigma_\varepsilon^4\sum_{i=1}^{N}\vartheta_i^2 > 0 \end{aligned}$$

This results is summarized in the following proposition.

Proposition 3.1 *For the price process given in expression (1), and in the absence of market microstructure noise, the realized variance measure is more efficient under BTS than under any other conceivable sampling scheme. Further, the efficiency gain associated with BTS, relative to CTS, increases with (i) an increase in the variability of trade intensity, and (ii) an increase in the variance of the price innovations.*

3.2 Presence of Market Microstructure Noise

We now turn to the case where market microstructure noise is present and start with the analysis of the restricted CPP-MA(1) model given by expression (1) with $q = 1$, $\rho_0 = 1$, and $\rho_1 = -1$. As we will see later on, this model is already sufficiently flexible to capture a lot of the market microstructure noise in actual data. Relevant moments for the CPP-MA(1) are given in expressions (9) through (11) in Appendix A (set $\rho = 0$). Now that price innovations are serially correlated, realized variance is a biased estimator of the integrated variance. Under GTS, this bias can be expressed as:

$$Bias\left(GTS_N\right) = E_\lambda\left[\sum_{i=1}^{N} R\left(t_i|\tau_i\right)^2\right] - \lambda_{(0,1)}\sigma_\varepsilon^2 = 2\sigma_\nu^2 \sum_{i=1}^{N}\left(1 - e^{-\lambda_i}\right) > 0$$

For CTS_N, the bias is given by the above expression with $\lambda_i = \int_{(i-1)\delta}^{i\delta} \lambda\left(u\right) du$ and $\delta = N^{-1}$. For BTS_N, the bias expression simplifies to $2N\sigma_\nu^2\left(1 - e^{-\lambda_N}\right)$. Note that when $N = 1$, all schemes coincide so that the bias is identical. What is less obvious is that when $N \to \infty$ all sampling schemes coincide as well. This is a consequence of the price process being discontinuous and of bounded variation: when $N \to \infty$ the timing and magnitude of *all* price changes will be observed regardless of the sampling scheme. Hence, the bias will be the same and equal to:

$$\lim_{N\to\infty} Bias\left(GTS_N\right) = \lim_{N\to\infty} 2\sigma_\nu^2 \sum_{i=1}^{N}\left(1 - e^{-\lambda_i}\right) = \lim_{N\to\infty} 2\sigma_\nu^2 \sum_{i=1}^{N} \lambda_i = 2\sigma_\nu^2 \lambda_{(0,1)}$$

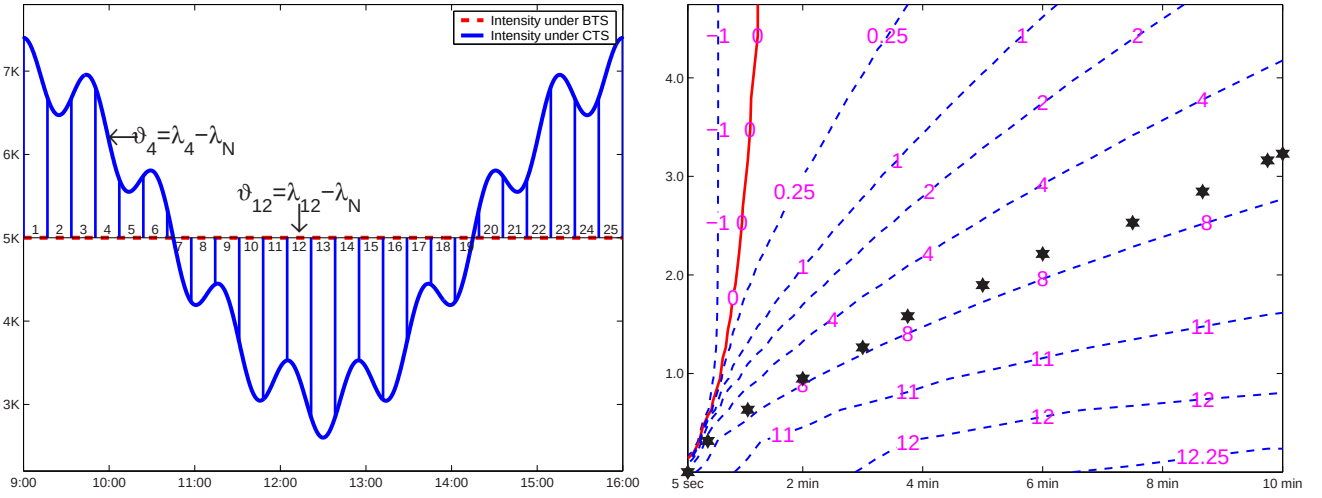
For all intermediate cases, i.e. $1 < N < \infty$, it is of interest to compare the relative magnitude of the bias under different sampling schemes:

$$Bias\left(GTS_N\right) - Bias\left(BTS_N\right) = 2e^{-\lambda_N}\sigma_\nu^2 \sum_{i=1}^{N}\left(1 - e^{-\vartheta_i}\right) < 0.$$

Some of these results are summarized in the following proposition:

Proposition 3.2 *For the price process given in expression (1), and in the presence of first order market microstructure noise, the bias of the realized variance measure under BTS is larger than under any other conceivable sampling scheme for* $1 < N < \infty$*. Further, the difference in bias increases with (i) an increase in the variability of trade intensity, and (ii) an increase in the variance of the market microstructure noise component.*

FIGURE 1: INEFFICIENCY OF CTS RELATIVE TO BTS



Note. Left panel: assumed trade intensity process over the trading day from 9:00 to 16:00 (average of 5000 trades per day). Right panel: iso curves of CTS loss (in percentages) as a function of the sampling frequency (horizonal axis) and noise ratio (vertical axis). The asterisks indicate the optimal sampling frequency for a given noise ratio and are the same under both BTS and CTS.

The above result suggests that the bias can be minimized by adopting a sampling scheme that is as much "different" from BTS as possible, i.e. set the length of the sampling intervals proportional to the integrated trade intensity leading one to sample (in) frequent when markets are (fast) slow. Even though this sampling strategy may reduce the bias, it may also have the undesirable side effect of increasing the variance of the estimator. It thus seems appropriate to consider the MSE for which an explicit expression can be derived based on the return moments listed in Appendix A. Using this expression (not reported to conserve space) and the model parameters, the optimal sampling frequency can then be determined by minimizing the MSE over N, i.e. the number of sampling points. We will come back to this in Section 4. Again, it is of interest to compare the relative magnitude of the MSE under GTS and BTS, i.e.

$$\begin{aligned} MSE\left(GTS\right) - MSE\left(BTS\right) \quad = \quad & 2\sigma_{\varepsilon}^{4}\sum_{i=1}^{N}\vartheta_{i}^{2} + 4e^{-\lambda_{N}}\sigma_{\nu}^{2}(\sigma_{\varepsilon}^{2}\lambda_{(0,1)} - 3\sigma_{\nu}^{2})\sum_{i=1}^{N}(e^{-\vartheta_{i}} - 1) \\ & + 4\sigma_{\nu}^{4}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\begin{array}{c}\left(1-e^{-\lambda_{i}}\right)\left(1-e^{-\lambda_{j}}\right)\left(2+e^{-\lambda_{i,j}}\right) \\ -\left(1-e^{-\lambda_{N}}\right)^{2}\left(2+e^{-(j-i-1)\lambda_{N}}\right)\end{array}\right) \\ & + 4\sigma_{\nu}^{2}\sigma_{\varepsilon}^{2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\begin{array}{c}2\lambda_{N}\left(e^{-\lambda_{N}}-1\right) \\ -\lambda_{i}\left(e^{-\lambda_{j}}-1\right) - \lambda_{j}\left(e^{-\lambda_{i}}-1\right)\end{array}\right) \end{aligned}$$

Unfortunately, we cannot analytically determine over which range of sampling frequencies the BTS scheme outperforms the CTS scheme in terms of MSE without further specification of the intensity process. We know that when $N = 1$ and $N \to \infty$ all sampling schemes coincide and hence the difference in MSE will be zero. Further, we know that the bias under BTS is strictly larger than under GTS for all intermediate cases. While we intuitively expect the variance of realized variance to be smaller under BTS, its magnitude relative to the bias

cannot be worked out straightforwardly. We therefore conduct a simulation exercise where we compute, for a given sampling frequency and noise ratio (i.e. $\sigma_\eta/\sigma_\varepsilon$), the percentage increase in MSE of realized variance under CTS relative to BTS. Throughout the remainder of this paper, we refer to this quantity as the "CTS loss". We specify the intensity process as a simple function of time, integrating to 5000 over the day (i.e. an expected 5000 trades per day) and set the variance of the efficient price innovation (σ_ε^2) set so that the annual return variance is equal to 25%. The left panel of Figure 1 plots the intensity process in calendar time (solid line) with the superimposed dashed line representing the intensity process on a business time scale as defined above. The right panel of Figure 1 plots iso-curves for the CTS loss with the sampling frequency on the horizonal axis ranging between 5 seconds and 10 minutes and the noise ratio on the vertical axis ranging between 0 and 4.5 (the empirical analysis in the next section computes an average optimal sampling frequency of 3 minutes and a noise ratio of 1.5 for IBM transaction data over the period Jan 2000 to Aug 2003).

Consistent with the discussion above, the right panel of Figure 1 shows that when market microstructure noise is absent (i.e. $\sigma_\eta/\sigma_\varepsilon = 0$), BTS outperforms CTS at any given sampling frequency. On the other hand, when noise is introduced the MSE reduction under BTS diminishes and can even turn negative. This occurs when the noise ratio / sampling frequency combination lies in the region to the left of the solid line. Additional simulations (not reported) suggest that an increase in the level of the intensity process causes the solid line to shift leftwards thereby reducing the area where CTS loss is negative while an increase in the variability of the intensity process leads to greater benefits associated with BTS to the right of the solid line. Ultimately, however, these results say little about the performance of BTS in practice unless we specify a sampling frequency. We therefore indicate the optimal sampling frequency under BTS, for given noise ratio, with an asterisk in the graph as this will be the frequency most relevant in empirical applications (the optimal sampling frequencies under CTS are virtually identical). Evidently, BTS is superior at and around the optimal sampling frequency. Moreover we find that this result is robust to alternative choice of model parameters and intensity processes.

An issue which remains open at this point is whether, and if so how, the above results are expected to change when the order of the MA process is raised. In particular, what will happen to the MSE and optimal sampling frequency in such a case? And also, will the properties of the BTS scheme change? To answer the first question, we investigate how the optimal sampling frequency and magnitude of the MSE changes when we move from an MA(1) specification to an MA(2). In doing so, we distinguish among two cases namely, (i) change ρ while keeping σ_ν^2 constant and (ii) change ρ while keeping σ_η^2 constant. The first case corresponds to *adding* noise and we find that this leads to an increase in MSE and a decrease of optimal sampling frequency. The second case corresponds to *altering the structure* of the noise dependence in which case the results can go either way. Both these finding are quite intuitive and are robust to the specific choice of parameter values and the order of the MA process. To answer the second question, we redo the above analysis for the MA(2) case and find that also these results are robust to higher order dependence.

4 Market Microstructure Noise in Practice: IBM Transaction Data

In this section we will discuss how the above methodology can be used in practice to (i) determine the optimal sampling frequency and (ii) measure the improvement in MSE resulting from BTS relative to CTS. The analysis will be based on the restricted CPP-MA(1) and CPP-MA(2) specifications using transaction level data for IBM. Our main findings are as follows. First, the estimation of the model is straightforward, delivers sensible results, and the estimated optimal sampling frequency is not biased by the measurement error in the model parameters. Second, the results for the CPP-MA(1) and CPP-MA(2) are very similar suggesting that, for our purposes here, there is no need to include higher order dependence in the noise component. Third, the optimal sampling frequency for IBM transaction data has decreased from 5 minutes in 2000 to 2.5 minutes in 2003. There is considerable day-to-day variation in the optimal sampling frequency which can be explained almost perfectly by changes in the noise ratio and market activity. Fourth, BTS outperforms CTS for each day in the sample with the MSE reduction under BTS being largest on days with irregular trading patterns.

4.1 The Data

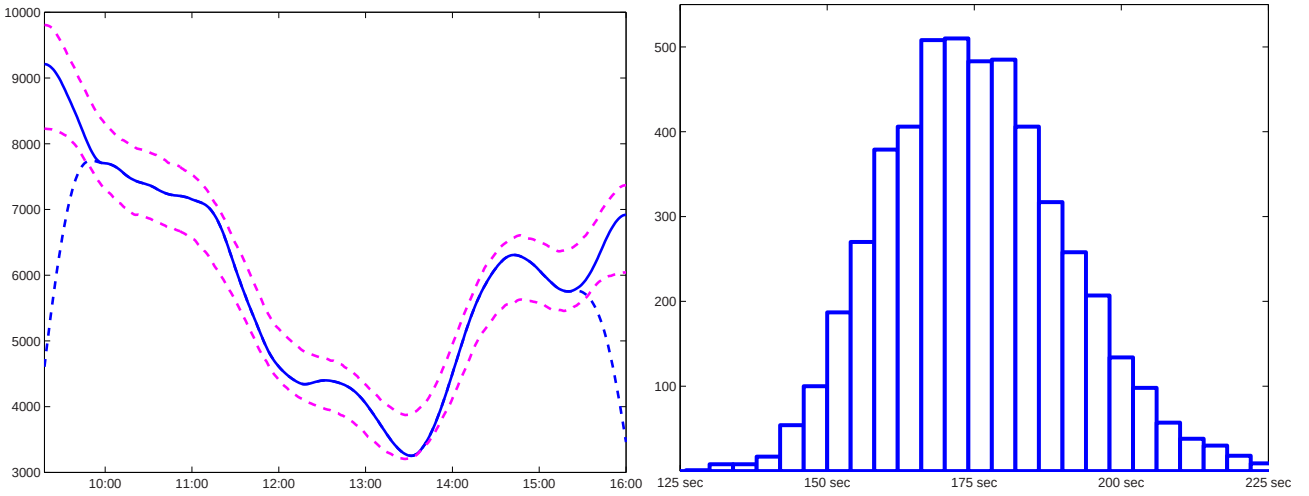
We use TAQ transaction data for IBM over the period from January 1, 2000 until 31 August 2003 (917 days). Data from all exchanges are considered but we discard transactions which take place before 9.45 and after 16.00. We also remove transactions with an error code. Outliers are dealt with by filtering for instantaneous price reversals. Let P_k denote the logarithmic price at which the k^{th} transaction is executed. For P_k to be removed by the filter, the following conditions need to be satisfied: $|R_k| = |P_k - P_{k-1}| > c$ and $R_{k+1} \in [-(1-w)R_k, -(1+w)R_k]$ for $0 < w < 1$. The parameter c controls the minimum magnitude of the "candidate outlier", while w specifies the region to which the subsequent transaction price needs to revert back to in order for it to be considered a price reversal. Based on some experimentation, we set $w = 0.25$ and c equal to 8 times a robust (interquartile) volatility estimate of transaction returns. With this specification of the filter, we find it successfully removes a total of 1358 "visibly significant" outliers. The final data set contains a total of 5,522,929 transactions.

4.2 Estimation of the Model Parameters

The restricted CPP-MA(2) model is given by expression (1) with $q = 2$, $\rho_0 = 1$, $\rho_1 = \rho - 1$, and $\rho_2 = -\rho$. For simplicity we assume that μ_ε and μ_ν are zero, which leaves us with the task of estimating $\left\{\sigma_\varepsilon^2, \sigma_\nu^2, \rho\right\}$ together with the intensity process $\lambda(t)$. Based on the variance and first and second order serial covariance of transaction returns on a given day we obtain estimates for the model parameters by solving the following three equations:

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

FIGURE 2: INTENSITY PROCESS & IMPACT OF MEASUREMENT ERROR



Note. Left panel: a non parametric bias-corrected estimate of the arrival intensity for IBM transactions on August 25, 2003 (solid line) with 1% and 99% bootstrapped confidence bounds (dashed lines). The downward sloping dashed lines near the edges represent the intensity estimate without bias correction. Right panel: histogram of optimal sampling frequency estimates in the presence of measurement error. The true optimal sampling frequency is 175 seconds.

and select the solution where $\sigma_\nu > 0$ and $|\rho| < 1$. Regarding the intensity process estimation, we note that the total number of transactions on a given day provides us with an unbiased estimator of the integrated intensity process. An estimate for $\lambda(t)$ can be obtained using standard non-parametric smoothing techniques for intensity estimation. As is the case for any non-parametric method, a choice of bandwidth (primary importance) and kernel (secondary importance) must be made. Based on extensive simulations for representative sample sizes and intensity processes, we found that a quartic kernel with a bandwidth of 0.10 days gives satisfactory results. This particular choice of bandwidth implies that for the estimation of the intensity at time t, we smooth transaction arrivals over the interval $[t - 37.5\text{min}, t + 37.5\text{min}]$. Since, we perform the estimation for each day in the sample separately, the estimator will be biased downwards at the edges (no data before open or after close). To resolve this we implement a "mirror image adjustment" discussed in Diggle and Marron (1988). Simulation results (not reported) indicate that this method works very well for the sample sizes we work with. As an illustration, the left panel of Figure 2 contains an estimate of the trading intensity process for IBM on August 25, 2003. The solid line is the bias adjusted estimator with the downward sloping dashed lines representing the unadjusted estimator close to the edges. To gauge the significance of the estimates, we compute bootstrapped confidence bounds (1% and 99%) based on the method developed in Cowling, Hall, and Phillips (1996) using 1000 replications. The resulting bounds (dashed lines) suggest that the intensity process can be estimated accurately from the data at hand, which is important since the computation of the efficiency gain of BTS over CTS crucially relies on this.

Since part of the focuss in the empirical part will be on estimating the optimal sampling frequency, it is important to study the impact of measurement error. Recall that the optimal sampling frequency is determined by minimizing the MSE expression of realized variance over N. Because the appropriate MSE expression is

Table 1: CPP-MA Estimation Results for IBM Transaction Data 2000-2003

	CPP-MA(1)					CPP-MA(2)				
	2000	2001	2002	2003	2000-2003	2000	2001	2002	2003	2000-2003
Noise Ratio										
median	1.876	1.339	1.630	1.353	**1.567**	1.745	1.336	1.753	1.512	**1.593**
min	0.945	0.604	0.703	0.787	0.604	0.996	0.915	0.973	0.804	0.804
max	4.232	2.548	6.929	2.603	6.929	4.404	3.517	4.891	2.751	4.891
Frequency										
median	315	176	190	144	**195**	289	179	208	162	**206**
min	55	42	65	73	42	87	98	102	68	68
max	934	461	1093	339	1093	769	772	796	346	796
Loss CTS										
median	3.620	2.564	2.882	3.348	**3.061**	3.698	2.537	2.837	3.360	**3.042**
min	0.682	0.607	0.702	0.991	0.607	0.712	0.459	0.537	0.991	0.459
max	40.24	50.91	38.56	36.37	50.91	40.37	55.93	37.74	36.36	55.93

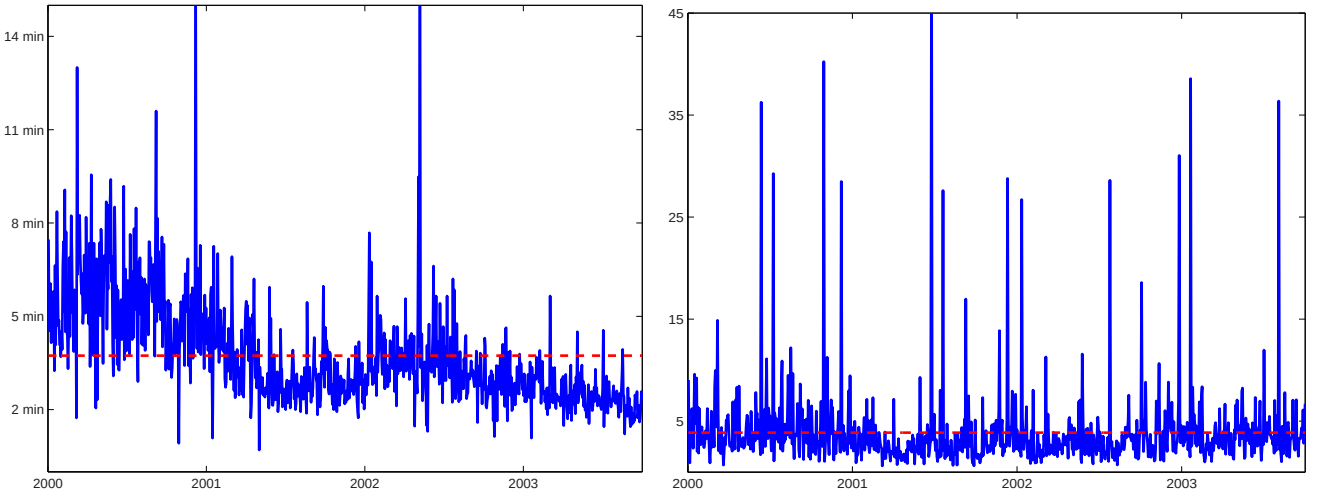
Note. This table contains summary statistics for the daily estimates of the noise ratio (upper panel), optimal sampling frequency in seconds (middle panel), and CTS loss in percentages (lower panel) for IBM transaction data over the period Jan 2000 until Aug 2003.

relatively complex, with the model parameters entering non-linearly, we undertake a simulation experiment to investigate the issue. Based on the CPP-MA(2) model and realistic parameter values (i.e. $\rho = 0.6$, $\lambda_{(0,1)} = 5000$, $\sigma_\nu/\sigma_\varepsilon = 1.1$, annualized return volatility of 25%), we simulate a day worth of transaction data. We then estimate the model parameters, integrated intensity process, and optimal sampling frequency (under BTS), compare these to their actual values and repeat the procedure several times. The right panel of Figure 2 reports a histogram of estimated optimal sampling frequency in the presence of measurement error based on 5000 simulation runs. The optimal sampling frequency based on the true parameters is 175 seconds. As expected, the estimated sampling frequencies do deviate from their true values but, quite surprisingly, the measurement error does not lead to a bias in the estimates. This finding contrasts sharply to the results reported in Bandi and Russell (2003) which illustrate that, within their proposed framework, measurement error renders the estimated optimal sampling frequency severely biased.

4.3 Empirical Results

As mentioned above, we estimate the intensity process plus the restricted CPP-MA(1) and CPP-MA(2) model parameters for each day in the sample separately and then determine (i) the optimal sampling frequency under BTS by minimizing the analytic MSE expression over N and (ii) the percentage increase in MSE under CTS relative to BTS, i.e. the CTS loss, at the optimal BTS frequency. Figure 3 contains a time series of the estimated

FIGURE 3: OPTIMAL SAMPLING FREQUENCY AND INEFFICIENCY OF CTS



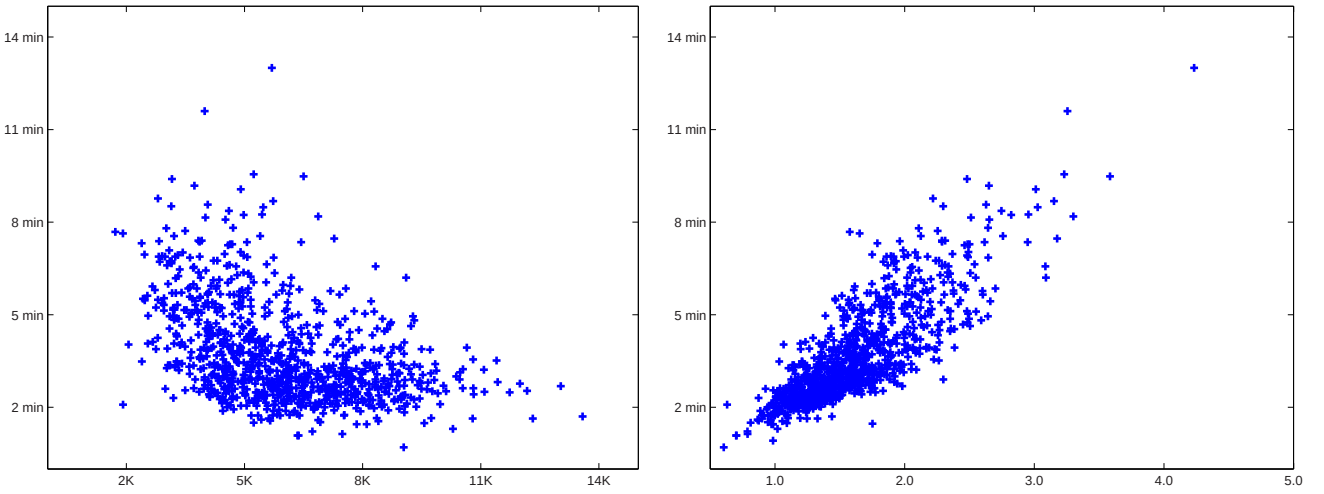
Note. The estimated daily optimal sampling frequency under BTS for IBM transaction data over the period Jan 2000 until Aug 2003 (left panel) with the associated CTS loss in percentage points (right panel).

optimal sampling frequency (left panel) and CTS loss (right) panel and Table 1 reports additional summary statistics.

Regarding the specification of the noise term, it is clear from Table 1 that the results are very similar for both models. The CPP-MA(2) model picks up a little more dependence which leads to a slightly lower optimal sampling frequency on average (206 seconds instead of 195 for the CPP-MA(1)) but the sample average of the daily estimates for ρ is still very close to zero (i.e. 0.0026). Hansen and Lunde (2004a) argue that the higher order non-iid type noise, such as the MA(2) specification, may be required for the modelling of the market microstructure in practice but that this is more important for the modelling of mid-quotes as opposed to transaction data. It therefore seems reasonable to conclude that the CPP-MA(1) is sufficiently flexible for the purpose and data set at hand. Although beyond the scope of the current paper, in future research we may explore extended versions of the CPP-MA model to investigate the importance of including higher order dependence, leverage effects, and non-Gaussian innovations.

The results reported in Figure 3 and Table 1 show that there is a downward trend in the optimal sampling frequency and considerable day-to-day variation in the estimates. For instance, the average estimated optimal sampling frequency over 2000 is 5 minutes compared to 2.5 minutes over the first 8 months of 2003. Also, on some days the optimal sampling frequency can be as low as 1 minute while on other days it may exceed 10 minutes. We recognize that some of this variation in optimal sampling frequency may be due to measurement error in the daily estimates of the model parameters. However, the simulation results reported above indicate that the impact of the measurement error does not come close to explaining the observed variation in optimal sampling frequency. It is therefore clear that the optimal sampling frequency *does* vary substantially over time. Although this finding has important implications for the computation of realized variance in practice, commonly

FIGURE 4: DETERMINANTS OF THE OPTIMAL SAMPLING FREQUENCY



Note. Scatterplot of the estimated optimal sampling frequency under BTS versus the number of transactions (left panel) and noise ratio (right panel). The data is IBM transaction data over the period Jan 2000 until Aug 2003.

done at a fixed frequency for the entire sample, it is not a surprising results because market activity, market microstructure, and the statistical properties of returns, are known to change over time. To explore this point a little further, Figure 4 reports a scatter plot of the optimal sampling frequency in minutes on the vertical axis against the number of transactions (left panel) and noise ratio (right panel). While one would be tempted to argue that the price process can be sampled at higher frequencies on active trading days, this relation appears weak. On the other hand, variation in the optimal sampling frequency seems to be much more closely related to changes in the noise ratio. Another way to illustrate this issue is by means of the following linear regression:

$$\log \hat{F}_t = \omega + \beta_1 \log \hat{\sigma}_{\nu,t} + \beta_2 \log \hat{\sigma}_{\varepsilon,t} + \beta_3 \log \hat{\lambda}_{(t,t+1)} + \epsilon_t \tag{3}$$

where $\hat{F}_t$ denotes the estimated day-t optimal sampling frequency based on the CPP-MA(1) model and $\hat{\lambda}_{(t,t+1)}$ is equal to the total number of transactions on day-t as an unbiased estimator of the integrated intensity. Clearly, there is an error in variable problem which we ignore here. The regression results reported in Table 2 confirm our previous finding. In particular, less than 25% of the variation in the optimal sampling frequency is explained by changes in market activity while this increases dramatically to just under 70% for the noise ratio. Quite surprisingly, combining the two variables uncovers an almost perfect relationship between the optimal sampling frequency, noise ratio, and number of transactions. Although intuitive in that less noise, more signal, and more transactions lead to a higher sampling frequency, this relation is certainly not obvious from the derived analytic MSE expressions. It is often argued that one can sample more frequently in liquid markets than in illiquid ones (Bandi and Russell 2003). Turning this argument around, the uncovered relationship suggests that a possibly non-linear transformation of the noise ratio and the number of transactions on a given day could provide a new measure for market liquidity. This issue will be pursued in future research.

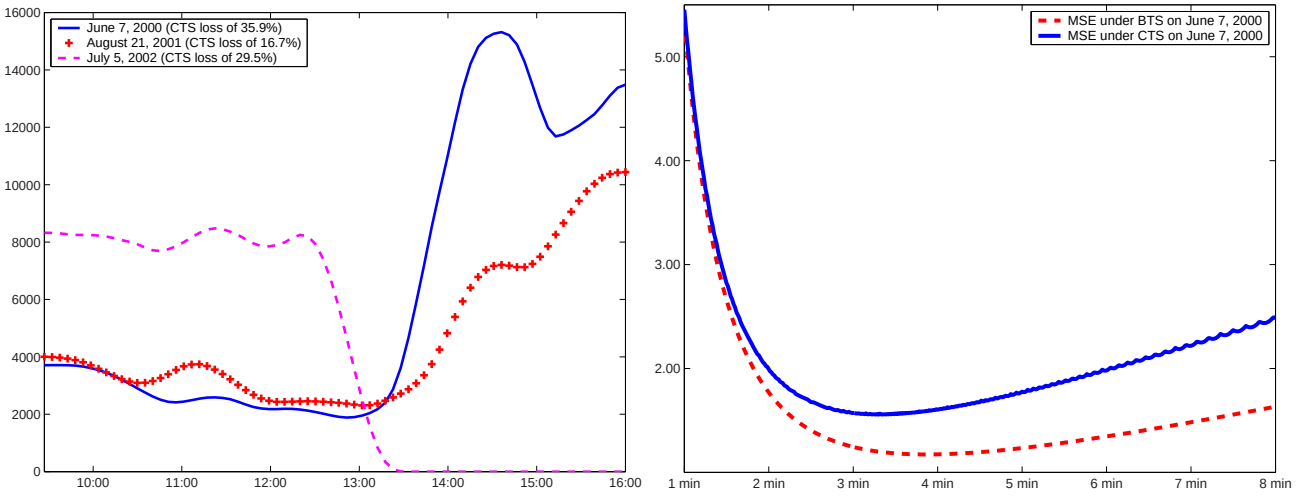
Table 2: Optimal Sampling Frequency Regression for IBM Transaction Data 2000-2003

Model	ω	β_1	β_2	β_3	R^2	LogLik	DW
I	10.22 (24.67)			-0.566 (12.00)	23.514%	-342.66	1.04
II	6.152 (17.79)	1.270 (32.5)	-1.107 (21.7)		69.494%	78.325	0.64
III	10.51 (5157)	1.340 (4491)	-1.340 (3432)	-0.6701 (3091)	99.998%	4528.6	1.89

Note. Regression results for $\log \hat{F}_t = \omega + \beta_1 \log \hat{\sigma}_{\nu,t} + \beta_2 \log \hat{\sigma}_{\varepsilon,t} + \beta_3 \log \hat{\lambda}_{(t,t+1)} + \epsilon_t$ where F denotes the day$-t$ estimated optimal sampling frequency under BTS based on the restricted CPP-MA(1) model. HAC consistent t-statistics are in parenthesis below the estimated coefficient. The column "DW" reports the Durbin Watson statistic. April 18, 2002 is taken out to remove an outlier in the optimal sampling frequency of 1093 seconds.

We are now left with the central question whether BTS does indeed lead to a lower MSE of realized variance. As shown above, the key to BTS superiority is the diurnal pattern of trade activity. Using the non-parametric techniques mentioned above, we estimate the trading intensity process for each day in the sample and then use the CPP-MA model to compute the daily difference in MSE under CTS relative to BTS at the optimal sampling frequency under BTS (whether to evaluate the MSE at the optimal sampling frequency under BTS or CTS here does not have a material impact on the results since both lie closely together in most cases). The results reported in Table 1 and the right panel of Figure 3 show that BTS did achieve a lower MSE than CTS on every single day in the sample. Although the average increase in MSE under CTS is a modest 3%, the most important results is that BTS never does worse. A closer look at the instances when the CTS loss exceeded 15% indicates that the gains of using the BTS scheme are particularly large on days with highly irregular trading patterns, early market closures, or sudden moves in market activity. The left panel of Figure 5 plots the trading intensity estimates for three of such days. For instance, on June 7, 2000 IBM trading activity surged due to a favorable analyst report. On that day the Dow Jones Business News headlined "Wall Street Closes Higher, Paced By IBM Rebound On Goldman Sachs Comments" and reported that "A late-day rally in IBM shares helped push stocks higher Wednesday…International Business Machines (IBM) jumped 8 3/8 to 120 3/4 after Goldman Sachs analyst Laura Conigliaro told CNBC that the computer maker should see revenue improvements in the second half of the year" (source: Factiva). The reason for the trading pattern on August 21, 2001 is unclear while on July 5, 2002 the market closed at 13.00. Despite the unusual market activity, the estimated CPP-MA(2) parameters for these days are very much in line with those on other days suggesting that this cannot explain the gains associated with BTS (i.e. June 7, 2000: $\rho = 0.0308$, $\sigma_\nu/\sigma_\varepsilon = 1.8060$, $\lambda_{(0,1)} = 6274$, optimal sampling frequency 232 seconds, CTS loss of 35.9%, annualized daily return volatility 21.8%. August 21, 2001: $\rho = -0.0096$, $\sigma_\nu/\sigma_\varepsilon = 1.3404$, $\lambda_{(0,1)} = 4600$, optimal sampling frequency 192 seconds, CTS loss of 16.7%, annualized daily return volatility 14.2%. July 5, 2002: $\rho = -0.1773$, $\sigma_\nu/\sigma_\varepsilon = 1.2026$, $\lambda_{(0,1)} = 4282$, optimal sampling frequency 178 seconds, CTS loss of 29.5%, annualized daily return volatility 13.0%). The right panel of Figure 5 plots the MSE of

FIGURE 5: ESTIMATED TRADING INTENSITY AND MSE ON BTS FAVORABLE DAYS



Note. Left panel: non-parametric estimate of trading intensity. Right panel: MSE of realized variance under BTS and CTS for sampling frequencies between 1 and 8 minutes.

realized variance for June 7, 2000 based on the CPP-MA(2) model under CTS and BTS for sampling frequencies between 1 and 8 minutes. It can be seen that the reduction in MSE is substantial around the optimal sampling frequency. In fact, sampling at 8 minutes in BTS achieves the same MSE as sampling at 3 minutes under CTS!

5 Conclusions

In this paper we have proposed a new framework in which to analyze the properties of realized variance in the presence of market microstructure effects. The logarithmic security price is modelled semi-parametrically as a compound Poisson process which allows for a general MA(q) dependence structure on price innovations and leaves the intensity process unspecified. The conditional characteristic function of the price process is derived, based on which we investigate the bias and MSE properties of realized variance both at different sampling frequencies and different sampling schemes. The main message of this paper is that business time sampling is superior to the common practice of calendar time sampling in that it reduces the MSE of realized variance when returns are sampled at or around the optimal sampling frequency. The empirical analysis confirms this finding.

A Appendix

Proof of Theorem 2.1. For a general $MA(q)$-innovation structure define

$$r_t(h) = \sum_{j=0}^{h-1}(\varepsilon_{t-j} + \eta_{t-j}) = \sum_{j=0}^{h-1}\varepsilon_{t-j} + \sum_{j=0}^{h-1}\sum_{i=0}^{q}\rho_i\nu_{t-j-i} = \sum_{j=0}^{h-1}\varepsilon_{t-j} + \sum_{j=0}^{h+q-1}\sum_{i=\max(0,j-h+1)}^{\min(j,q)}\rho_i\nu_{t-j}$$

for $h > 0$ and $r_t(0) = 0$. From this it can be seen that the mean return $E[r_t(h)] = h(\mu_\varepsilon + \mu_v\overline{\rho})$ where $\overline{\rho} = \sum_{i=0}^{q}\rho_i$ and the variance $V[r_t(h)] \equiv \Sigma_q(h)$ is equal to:

$$\begin{aligned}\Sigma_q(h) &= h\sigma_\varepsilon^2 + h\sigma_\nu^2\sum_{i=0}^{q}\rho_i^2 + 2\sigma_\nu^2\sum_{j=1}^{\min(q,h)}\sum_{i=j}^{q}(h-j)\rho_i\rho_{i-j} \\ &= h\left(\sigma_\varepsilon^2 + \sigma_\nu^2\overline{\rho}^2\right) - 2\sigma_\nu^2\sum_{j=1}^{q}\sum_{i=j}^{q}j\rho_i\rho_{i-j} \qquad \text{for } h \geq q \end{aligned} \tag{4}$$

Moreover, the covariance between $r_t(h)$ and $r_{t+k+m}(k)$ for $q > m \geq 0$ can then be worked out as:

$$\begin{aligned}\Omega_q(h,k,m) &\equiv Cov\left(r_t(h), r_{t+k+m}(k)\right) \\ &= Cov\left(\sum_{j=0}^{h+q-1}\sum_{i=\max(0,j-h+1)}^{\min(j,q)}\rho_i\nu_{t-j}, \sum_{j=0}^{k+q-1}\sum_{i=\max(0,j-k+1)}^{\min(j,q)}\rho_i\nu_{t-(j-k-m)}\right) \\ &= Cov\left(\sum_{j=0}^{h+q-1}\sum_{i=\max(0,j-h+1)}^{\min(j,q)}\rho_i\nu_{t-j}, \sum_{w=-k-m}^{q-m-1}\sum_{i=\max(0,w+m+1)}^{\min(w+k+m,q)}\rho_i\nu_{t-w}\right) \\ &= \sigma_\nu^2\sum_{b=0}^{\min(h+q-1,q-m-1)}\left(\sum_{i=\max(0,b-h+1)}^{\min(b,q)}\rho_i\right)\left(\sum_{i=\max(0,b+m+1)}^{\min(b+k+m,q)}\rho_i\right)\end{aligned} \tag{5}$$

and $\Omega_q(h,k,m) = 0$ for $q \leq m$. Due to the joint normality of $r_t(h)$ and $r_{t+k+m}(k)$, their characteristic function can be expressed as:

$$\begin{aligned}\phi_q(h,k,m) &\equiv E\left(\exp\left\{i\xi_1 r_t(h) + i\xi_2 r_{t+k+m}(k)\right\}\right) \\ &= \exp\left\{i\left(\mu_\varepsilon + \mu_v\overline{\rho}\right)(\xi_1 h + \xi_2 k) - \frac{1}{2}\left(\xi_1^2\Sigma_q(h) + \xi_2^2\Sigma_q(k) + 2\xi_1\xi_2\Omega_q(h,k,m)\right)\right\}\end{aligned} \tag{6}$$

Using the above, the joint characteristic function of returns, $R(t_1|\tau_1)$ and $R(t_2|\tau_2)$ for $t_2 \geq t_1 + \tau_2$, conditional on the intensity process, can then be written as:

$$\begin{aligned}\phi_{RR}(\xi_1,\xi_2|t_1,t_2,\tau_1,\tau_2,q) &\equiv E_\lambda\left(\exp\left\{i\xi_1 R(t_1|\tau_1) + i\xi_2 R(t_2|\tau_2)\right\}\right) \\ &= \sum_{m=0}^{\infty}\sum_{h=0}^{\infty}\sum_{k=0}^{\infty}\phi_q(h,k,m)\frac{\lambda_2^k}{k!e^{\lambda_2}}\frac{\lambda_1^h}{h!e^{\lambda_1}}\frac{\lambda_{1,2}^m}{m!e^{\lambda_{1,2}}}.\end{aligned}$$

with $\lambda_j, \lambda_{i,j}$ as defined in expression (2). This infinite triple summation over m, h, and k from zero to infinity can be simplified somewhat by decomposing it as follows:

$$\sum_{m=0}^{\infty}\sum_{h=0}^{0}\sum_{k=0}^{0} + \sum_{m=0}^{\infty}\sum_{h=q}^{\infty}\sum_{k=q}^{\infty} + \sum_{m=0}^{\infty}\sum_{h=1}^{q-1}\sum_{k=1}^{\infty} + \sum_{m=0}^{\infty}\sum_{h=q}^{\infty}\sum_{k=1}^{q-1} + \sum_{m=0}^{\infty}\sum_{h=q}^{\infty} + \sum_{m=0}^{\infty}\sum_{k=q}^{\infty} + \sum_{m=0}^{\infty}\sum_{h=1}^{q-1} + \sum_{m=0}^{\infty}\sum_{k=1}^{q-1}$$

where the double summations have the missing third argument set to zero.

$$\phi_q(h,k,m) = \begin{cases} \exp\left\{i\xi_1 h(\mu_\varepsilon + \mu_v\overline{\rho}) - \frac{1}{2}\xi_1^2\Sigma_q(h)\right\} & \text{for } h<q; k=0 \\ \exp\left\{\xi_1^2 c_0 + hc_1\right\} & \text{for } h\geq q; k=0 \\ \exp\left\{i(\xi_1 h + i\xi_2 k)(\mu_\varepsilon + \mu_v\overline{\rho}) - \frac{1}{2}\xi_1^2\Sigma_q(h) - \frac{1}{2}\xi_2^2\Sigma_q(k) - \xi_1\xi_2\Omega_q(h,k,m)\right\} & \text{for } h<q; k<q \\ \exp\left\{i\xi_1 h(\mu_\varepsilon + \mu_v\overline{\rho}) + \xi_2^2 c_0 + kc_2 - \frac{1}{2}\xi_1^2\Sigma_q(h) - \xi_1\xi_2\Omega_q(h,k,m)\right\} & \text{for } h<q; k\geq q \\ \exp\left\{i\xi_2 k(\mu_\varepsilon + \mu_v\overline{\rho}) + \xi_1^2 c_0 + hc_1 - \frac{1}{2}\xi_2^2\Sigma_q(k) - \xi_1\xi_2\Omega_q(h,k,m)\right\} & \text{for } h\geq q; k<q \\ \exp\left\{\left(\xi_1^2+\xi_2^2\right)c_0 + hc_1 + kc_2 - \xi_1\xi_2\Omega_q(m)\right\} & \text{for } h\geq q; k\geq q \end{cases}$$

where we use the notation $\Omega_q(m) = \Omega_q(h\geq q, k\geq q, m)$ and

$$c_0 = \sigma_\nu^2 \sum_{j=1}^{q}\sum_{i=j}^{q} j\rho_i\rho_{i-j} \qquad \text{and} \qquad c_z = i\xi_z(\mu_\varepsilon + \mu_v\overline{\rho}) - \frac{1}{2}\xi_z^2\left(\sigma_\varepsilon^2 + \sigma_\nu^2\overline{\rho}^2\right) \quad \text{for } z\in\{1,2\}. \tag{7}$$

Some of the summations can now be simplified substantially, i.e. $\sum_{m=0}^{\infty}\sum_{h=1}^{q-1} \Rightarrow \sum_{h=1}^{q-1}$, $\sum_{m=0}^{\infty}\sum_{h=0}^{0}\sum_{k=0}^{0} \Rightarrow e^{-\lambda_1-\lambda_2}$, and

$$\begin{aligned} \sum_{m=0}^{\infty}\sum_{h=q}^{\infty} &\Rightarrow e^{-(\lambda_1+\lambda_2)}\sum_{h=q}^{\infty}\phi_q(h,0,0)\frac{\lambda_1^h}{h!} = \exp\left\{\xi_1^2 c_0 - \lambda_2 + \lambda_1\left(e^{c_1}-1\right)\right\}\left(1 - \frac{\Gamma(q,\lambda_1 e^{c_1})}{\Gamma(q)}\right) \\ \sum_{m=0}^{\infty}\sum_{k=q}^{\infty} &\Rightarrow e^{-(\lambda_1+\lambda_2)}\sum_{k=q}^{\infty}\phi_q(0,k,0)\frac{\lambda_2^k}{k!} = \exp\left\{\xi_2^2 c_0 - \lambda_1 + \lambda_2\left(e^{c_2}-1\right)\right\}\left(1 - \frac{\Gamma(q,\lambda_2 e^{c_2})}{\Gamma(q)}\right) \\ \sum_{m=0}^{\infty}\sum_{h=q}^{\infty}\sum_{k=q}^{\infty} &\Rightarrow \Psi_q(\xi_1,\xi_2,q,\Omega_q)\exp\left\{\left(\xi_1^2+\xi_2^2\right)c_0 + \lambda_1\left(e^{c_1}-1\right) + \lambda_2\left(e^{c_2}-1\right)\right\} \end{aligned}$$

where $\Gamma(w) = \int_0^\infty u^{w-1}e^{-u}du$ (gamma function), and $\Gamma(w,z) = \int_z^\infty u^{w-1}e^{-u}du$ (incomplete gamma function) and:

$$\Psi_q(\xi_1,\xi_2,q,\Omega_q) = \left(1 - \frac{\Gamma(q,\lambda_1 e^{c_1})}{\Gamma(q)}\right)\left(1 - \frac{\Gamma(q,\lambda_2 e^{c_2})}{\Gamma(q)}\right)\left(1 - \frac{\Gamma(q,\lambda_{1,2})}{\Gamma(q)} + \sum_{m=0}^{q-1} e^{-\xi_1\xi_2\Omega_q(m)}\frac{\lambda_{1,2}^m}{m!e^{\lambda_{1,2}}}\right) \tag{8}$$

Collecting all the yields the expression for $\phi_{RR}(\xi_1,\xi_2|t_1,t_2,\tau_1,\tau_2,q)$ given in Theorem 2.1. ■

Moment Expressions for the Restricted CPP-MA(2). Based on the characteristic function for the restricted CPP-MA(2), i.e. $\mu_\nu = \mu_\varepsilon = 0$, and $\rho_0 = 1, \rho_1 = \rho - 1$, and $\rho_2 = -\rho$, the following moment expressions can be derived:

$$E_\lambda\left[R(t_i|\tau_i)^2\right] = \lambda_i\sigma_\varepsilon^2 + 2\sigma_\nu^2\left(1+\rho^2\right)\left(1-e^{-\lambda_i}\right) - 2\sigma_\nu^2\rho\lambda_i e^{-\lambda_i} \tag{9}$$

$$\begin{aligned} E_\lambda\left[R(t_i|\tau_i)^4\right] &= 3\lambda_i\sigma_\varepsilon^4 + 3\lambda_i^2\sigma_\varepsilon^4 + 12\lambda_i\sigma_\varepsilon^2\sigma_\nu^2\left(1-\rho e^{-\lambda_i}+\rho^2\right) + 12\sigma_\nu^4\left(1+\rho^2\right)^2\left(1-e^{-\lambda_i}\right) \\ &\quad -12\sigma_\nu^4\rho\lambda_i e^{-\lambda_i}\left(2-\rho+2\rho^2\right) \end{aligned} \tag{10}$$

$$\begin{aligned} E_\lambda\left[R(t_i|\tau_i)^2 R(t_j|\tau_j)^2\right] &= \lambda_i\lambda_j\sigma_\varepsilon^4 + 2\sigma_\varepsilon^2\sigma_\nu^2\left(1+\rho^2\right)\left(\lambda_i+\lambda_j-\lambda_j e^{-\lambda_i}-\lambda_i e^{-\lambda_j}\right) \\ &\quad -2\lambda_i\lambda_j\rho\sigma_\varepsilon^2\sigma_\nu^2\left(e^{-\lambda_i}+e^{-\lambda_j}\right) - 4\sigma_\nu^4\left(1+\rho^2\right)^2\left(e^{-\lambda_j}-e^{-\lambda_j-\lambda_i}+e^{-\lambda_i}-1\right) \\ &\quad -4e^{-\lambda_i-\lambda_j}\sigma_\nu^4\rho\left(1+\rho^2\right)\left(\lambda_i e^{\lambda_j}-\lambda_i+\lambda_j e^{\lambda_i}-\lambda_j\right) + 4e^{-\lambda_i-\lambda_j}\sigma_\nu^4\rho^2\lambda_i\lambda_j \\ &\quad -2\sigma_\nu^4 e^{-\lambda_i-\lambda_{i,j}-\lambda_j}\left(\left(1+\rho^2\right)^2+\lambda_{i,j}\rho^2\right)\left(e^{\lambda_j}+e^{\lambda_i}-e^{\lambda_i+\lambda_j}-1\right) \\ &\quad +2\sigma_\nu^4 e^{-\lambda_i-\lambda_{i,j}-\lambda_j}\left(2\lambda_i\lambda_j\rho^2 - \rho\left(2-\rho+2\rho^2\right)\left(\lambda_i e^{\lambda_j}+\lambda_j e^{\lambda_i}-\lambda_j-\lambda_i\right)\right) \end{aligned} \tag{11}$$

References

Andersen, P. K., Ø. Borgan, R. Gill, and N. Keiding, 1993, *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

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, 2004a, *Continuous Time Approach to Financial Volatility*. Cambridge University Press, forthcoming.

Barndorff-Nielsen, O. E., and N. Shephard, 2004b, "Econometric Analysis of Realised Covariation: High Frequency Based Covariance, Regression and Correlation in Financial Economics," *forthcoming Econometrica*, 72.

Bowsher, C. G., 2002, "Modelling Security Market Events in Continuous Time: Intensity-Based, Multivariate Point Process Models," Manuscript Nuffield College, University of Oxford.

Carr, P., H. Geman, D. B. Madan, and M. Yor, 2002, "The Fine Structure of Asset Returns: An Empirical Investigation," *Journal of Business*, 75 (2), 305–332.

Carr, P., and L. Wu, 2003, "Time-Changed Lévy Processes and Option Pricing," forthcoming *Journal of Financial Economics*.

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.

Cox, J. C., and S. A. Ross, 1976, "The Valuation of Options for Alternative Stochastic Processes," *Journal of Financial Economics*, 3 (1/2), 145–166.

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

Geman, H., D. B. Madan, and M. Yor, 2001, "Time Changes for Levy Processes," *Mathematical Finance*, 11 (1), 79–96.

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.

Karlin, S., and H. M. Taylor, 1981, *A Second Course in Stochastic Processes*. Academic Press, New York.

Maheu, J. M., and T. H. McCurdy, 2003, "Modeling Foreign Exchange Rates with Jumps," manuscript University of Toronto.

——— , 2004, "News Arrival, Jump Dynamics and Volatility Components for Individual Stock Returns," *Journal of Finance*, 59 (2).

Meddahi, N., 2002, "A Theorical Comparison Between Integrated and Realized Volatility," *Journal of Applied Econometrics*, 17, 479–508.

Mürmann, A., 2001, "Pricing Catastrophe Insurance Derivatives," Manuscript Insurance and Risk Management Department, The Wharton School.

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

——— , 1968, "A Modified Compound Poisson Process with Normal Compounding," *Journal of the American Statistical Association*, 63, 607–613.

Rogers, L., and O. Zane, 1998, "Designing and Estimating Models of High-Frequency Data," Manuscript University of Bath.

Rydberg, T. H., and N. Shephard, 2003, "Dynamics of Trade-by-Trade Price Movements: Decomposition and Models," *Journal of Financial Econometrics*, 1 (1), 2–25.

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.