Realized Volatility When Sampling Times are Possibly Endogenous *

Yingying Li

Hong Kong University of Science and Technology

Per A. Mykland

University of Chicago and Oxford University

Eric Renault

University of North Carolina at Chapel Hill

Lan Zhang

University of Illinois at Chicago and Oxford University

Xinghua Zheng

Hong Kong University of Science and Technology

This version: December 1, 2009

*We are grateful to Andrew Patton and Neil Shephard, and the participants of the Stevanovich Center - CREATES 2009 conference for their comments and suggestions. Financial support from the Bendheim Center for Finance at Princeton University and the ISOM Department at HKUST (Li), the National Science Foundation under grants DMS 06-04758 and SES 06-31605 (Mykland and Zhang), and the University of British Columbia (Zheng) is also gratefully acknowledged. Address correspondence to: Xinghua Zheng, Department of ISOM, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong; (852) 2358 7750 or email: xhzheng@ust.hk.

Abstract

When estimating integrated volatilities based on high-frequency data, simplifying assumptions are usually imposed on the relationship between the observation times and the price process. In this paper, we establish a central limit theorem for the Realized Volatility in a general endogenous time setting. We also document that this endogeneity is present in financial data.

KEYWORDS: bias-correction, continuous semimartingale, discrete observation, efficiency, endogeneity, Itô process, realized volatility, stable convergence.

JEL CODES: C02; C12; C13; C14; C15; C22

1 Introduction

An important development in financial econometrics has been an asymptotic approach for inference on integrated (squared) volatility as estimated by realized variance. Substantial progress has been made on infill asymptotic theory to take advantage of the increasing availability of high frequency data. The earlier results in this direction were in probability theory (Jacod (1994), Jacod and Protter (1998)) while Barndorff-Nielsen and Shephard (2001, 2002) have been path-breaking for introducing this theory in econometrics. To be specific, the relevant asymptotic theory is based on two convergence results for an Ito process $dX_t = \mu_t dt + \sigma_t dW_t$ (with W_t Wiener process) observed at times $t_{n,i},\ i = 0, 1, \ldots .n$. The process X_t must be understood as a log-price so that $X_{t_{n,i}} - X_{t_{n,i-1}}$ is the continuously compounded rate of return over the corresponding time interval. The state of knowledge regarding asymptotic behavior of realized variance of high-frequency returns is then twofold.

First, if the observation times $t_{n,i}$ are stopping times such that the mesh of the partition $\max_i |t_{n,i} - t_{n,i-1}|$ goes to zero in probability, the realized variance $[X,X]_T = \sum_{t_{n,i}\leq T}(X_{t_{n,i}} - X_{t_{n,i-1}})^2$ is a consistent estimator of the quadratic variation $\langle X,X\rangle_T = \int_0^T \sigma_s^2 ds$.

Second, under some assumptions on the generating process of the times $t_{n,i}$ (see Mykland and Zhang (2006)), namely, if the so-called "quadratic variation of time" processes converges,

$$\lim_{n\to\infty} n \sum_{t_{n,i}\leq t} (t_{n,i} - t_{n,i-1})^2 = H_t, \tag{1}$$

where H_t is an adapted process, and the times $t_{n,i}$'s are independent of the X process, then $n^{1/2}([X,X]_T - \langle X,X\rangle_T)$ is asymptotically a mixture of normals whose mixture component is the variance coefficient equal to $2\int_0^T \sigma_s^4\, dH_s$, and is consistently estimated by $\frac{2n}{3}[X,X,X,X]_T$ where $\frac{n}{3}[X,X,X,X]_T = \frac{n}{3}\sum_{t_{n,i}\leq T}(X_{t_{n,i}} - X_{t_{n,i-1}})^4$ is the so-called

quarticity (Barndorff-Nielsen and Shephard (2002)). In the equidistant case, i.e., when $t_{n,i} = i/n$, (1) holds with $H_t = t$.

The equidistant case can also be generalized by using "time change" (Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008)). This induces some degree of endogeneity in the times, but not enough to induce the kind of bias we shall discuss here. Further generalizations of random times are given by Hayashi, Jacod, and Yoshida (2008) and Phillips and Yu (2007), but also when there is no asymptotic bias.

A striking feature of these results is that $[X, X, X]_T = \sum_{t_{n,i} \leq T} (X_{t_{n,i}} - X_{t_{n,i-1}})^3$ never comes into the picture. The key reason for that is that, even when conveniently scaled by $n^{1/2}$, this quantity generally vanishes asymptotically. To see this, first note that with constant volatility $\sigma_t = \sigma$, $\mu_t = 0$, and regular deterministic sampling $t_{n,i} = \frac{i}{n}$, we have:

$$n^{1/2}[X, X, X]_T = n^{1/2}\sigma^3 \sum_{t_{n,i} \leq T} (W_{t_{n,i}} - W_{t_{n,i-1}})^3 =_{\mathcal{L}} \sigma^3 \frac{T^{3/2}}{n} \sum_{i=1}^{n} U_i^3,$$

where the U_i are i.i.d. standard normal. Thus, by the law of large numbers:

$$\lim_{n \to \infty} n^{1/2}[X, X, X]_T = 0 \tag{2}$$

By a standard predictability argument, the property (2) remains clearly true when considering a stochastic volatility process σ_t in the context of regular deterministic sampling. It is in particular worth stressing that the well-documented skewness in stock returns as introduced by leverage effect (non-zero instantaneous correlation between σ_t and W_t) does not bring a non-zero limit for $n^{1/2}[X, X, X]_T$. Since stochastic volatility can be subsumed into a random time change, this remark also implies that even random

sampling times drawn according to a fixed random time change (see e.g. Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008)) will not destroy the result (2). The same applies to the result of Hayashi, Jacod, and Yoshida (2008) and Phillips and Yu (2007).

The focus of interest of this paper is a situation in which endogeneity of times does matter because it implies a non-zero limit for $n^{1/2}[X,X,X]_T$. The main theoretical result is that, in such circumstances, for the normalized error $n^{1/2}([X,X]_T - \langle X,X\rangle_T)$, the asymptotic Mean-Squared-Error (MSE) is still equal to $\lim_n \frac{2n}{3}[X,X,X,X]_T$ (which coincides with the asymptotic variances reported by the earlier papers), but must be decomposed differently. We will have instead a bias term which is non-zero if and only if the limit in (2) is no longer zero. The remaining term is the variance of a normal distribution.

Consistently estimating the aforementioned bias and variance should allow taking advantage of the informational content of endogenous sampling times to improve upon the common accuracy of volatility estimators. While a similar issue had already been addressed by Duffie and Glynn (2004) and Aït-Sahalia and Mykland (2003) (resp. Renault and Werker (2009)) in a parametric (resp. semi-parametric) context, this paper is the first to propose a model free approach. A related result has just been arrived at, independently and concurrently, by Fukasawa (2009), but with a substantially more opaque theoretical development.

On the empirical side, the paper shows that this endogeneity of time is actually present in the financial data. We use a large set of days for providing compelling evidence that the daily quantity $\lim_{n\to\infty} n^{1/2}[X,X,X]_T$ is not zero. $\lim_{n\to\infty} n^{1/2}[X,X,X]_T$ can actually be interpreted in terms of a measure of correlation between volatility and time. We also provide empirical evidence that the quarticity does not have the previously reported forms.

As extensively discussed by Renault and Werker (2009), a model-free measurement of the significant correlation between volatility and duration (between transactions or quote changes) is important both for economic theory of financial markets and for further developments on the estimation of continuous time processes in finance. Statistical evidence that this correlation is actually negative confirms the common wisdom that more news coming into the markets will simultaneously bring more volatility and more frequent transactions or quote changes. The mere fact that this correlation is not zero implies that a diffusion model observed with such random times ought not be estimated by simply plugging the random dates into the diffusion transition density function. Even a discrete time GARCH model with random time stamps should take this correlation into account by contrast with the currently available models (Grammig and Wellner (2002), Meddahi, Renault, and Werker (2006)). The continuous time framework should actually help to provide structural underpinnings to the GARCH approach to high frequency data proposed by Engle (2000).

The main theorem on the resulting new decomposition of the asymptotic mean squared error for quadratic variation estimation is developed in Section 2. This is done in the simplest case without microstructure noise. Theoretical illustrations are provided in Section 3, and tests for non-nullity of the endogeneity of times are devised in Section 4, with empirical results in Section 5. A simulation study is carried out in Section 6. The proof of the main theorem is in the Appendix.

2 Main Result

We use the usual Itô process model

$$dX_t = \mu_t dt + \sigma_t dW_t, \tag{3}$$

where W_t is a Wiener process. The target of inference is

$$\langle X, X \rangle_t = \int_0^t \sigma_s^2 ds. \tag{4}$$

DEFINITION 1. *(Stable Convergence.) Suppose that X_t, μ_t, and σ_t are adapted to filtration $(\mathcal{F}_t)$. Let Z_n be a sequence of $\mathcal{F}_T$-measurable random variables, We say that Z_n converges stably in law to Z as $n \to \infty$ if Z is measurable with respect to an extension of $\mathcal{F}_T$ so that for all $A \in \mathcal{F}_T$ and for all bounded continuous g, $EI_A g(Z_n) \to EI_A g(Z)$ as $n \to \infty$.*

For further discussion of stable convergence, see Rényi (1963), Aldous and Eagleson (1978), Chapter 3 (p. 56) of Hall and Heyde (1980), Rootzén (1980) and Section 2 (p. 169-170) of Jacod and Protter (1998).

THEOREM 1. *Let μ_t and σ_t^2 be adapted to a filtration $(\mathcal{F}_t)$, integrable, and locally bounded, and that $\sigma_t^2 \geq c > 0$, where c is nonrandom. Also assume that for some $\epsilon > 0$,*

$$\max |t_{n,i+1} - t_{n,i}| = o_p(n^{-(\frac{2}{3}+\epsilon)}). \tag{5}$$

Further assume that (for all t)

$$n[X, X, X, X]_t \xrightarrow{p} \int_0^t u_s \, ds \quad \textit{and} \tag{6}$$

$$n^{1/2}[X, X, X]_t \xrightarrow{p} \int_0^t v_s \, ds, \tag{7}$$

where u_t and $|v_t|$ are integrable. Finally, assume that the filtration $(\mathcal{F}_t)$ is generated by

finitely many continuous martingales. Then, stably in law:

$$n^{1/2}([X,X]_t - \langle X,X\rangle_t) \rightarrow \underbrace{\frac{2}{3}\int_0^t \frac{v_s}{\sigma_s^2}\, dX_s}_{asymptotic\ bias} + \int_0^t \sqrt{\frac{2}{3}u_s - \frac{4}{9}\frac{v_s^2}{\sigma_s^2}}\, dB_s,$$

where B_t is a Brownian-motion independent of the underlying σ-field.

It is worth interpreting this result in relation with the control variable approach in Monte Carlo estimation. When one wants to estimate the variance σ^2 of a random variable Z from an i.i.d. sample $Z_1, ...Z_n$, the naive estimator can in general be improved if one has the extra information that the expectation $E(Z)$ is zero. In this case, we know from the control variable principle that the unbiased estimator σ^{*2} of σ^2 with minimum (asymptotic) variance is the residual of the regression of the naive estimator on the sample mean of the Zs :

$$\sigma^{*2} = \frac{1}{n}\sum_{i=1}^{n} Z_i^2 - b^* \frac{1}{n}\sum_{i=1}^{n} Z_i$$

where b^* is the sample counterpart of the population regression coefficient:

$$b = \frac{Cov(Z^2, Z)}{Var(Z)} = \frac{E(Z^3)}{Var(Z)}$$

In other words, as soon as the variable Z has a non-zero skewness, the knowledge that it has a zero expectation allows us to improve the estimator in the sense of lowering its variance. This can be summarized by a limit result which is actually a particular case of theorem 1 above:

$$n^{1/2}[\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - \sigma^2] \longrightarrow V + W$$

where the convergence is in distribution and V and W are two independent normal variables defining the joint limit distribution of $\frac{b}{n^{1/2}}\sum_{i=1}^{n} Z_i$ and $n^{1/2}[\sigma^{*2} - \sigma^2]$. Coming from the naive estimator to σ^{*2}, the gain in asymptotic variance is equal to the variance of the normal variable V. Similarly, we can decompose the estimation error $n^{1/2}([X,X]_t - \langle X,X\rangle_t)$ in the following way. First, we note that it is asymptotically equivalent to $n^{1/2}M_t$ where M_t is the local martingale:

$$M_t = \sum_{t_{n,i}\leq t} (X_{t_{n,i}} - X_{t_{n,i-1}})^2 + (X_t - X_{t^*})^2 - \int_0^t \sigma_s^2 \, ds$$

and $t^* = \max\{t_{n,i}; t_{n,i} \leq t\}$. By Ito's lemma, this local martingale can be rewritten as:

$$M_t = 2\sum_{t_{n,i}\leq t} \int_{t_i}^{t_{i+1}} (X_s - X_{t_i}) dX_s + 2\int_{t^*}^{t} (X_s - X_{t^*}) \, dX_s$$

Following Mykland and Zhang (2006), we have a continuous time analog of the regression above by decomposing this local martingale as:

$$n^{1/2}M_t = \int_0^t g_s dX_s + M_t^{*(n)}, \text{ with } P\lim[M^{*(n)}, X]_t = 0$$

The process (g_t) solution of this equation is actually characterized by:

$$\int_0^t g_s\sigma_s^2 ds = P\lim \ \langle n^{1/2}M, X\rangle_t = \frac{2}{3}\int_0^t v_s \, ds$$

where the last equality, still a consequence of Ito's lemma, is explicitly derived in the Appendix. Hence:

$$g_s = \frac{2}{3}\frac{v_s}{\sigma_s^2}$$

and the so-called bias term in the theorem 1 above is nothing but the continuous-time regression $\int_0^t g_s dX_s$ of the naive estimation error on the process X itself. The control

variable principle still allows us an efficiency gain with respect to the naive estimator when the endogeneity of time produces a non-zero "continuous-time skewness" as manifested by a non-zero tricity. Note that this control variable works because the Girsanov theorem gives us the continuous time analog of the zero-expectation information above. More precisely, the continuous time regression of the estimation error on the process X itself does not produce any perverse bias because for the purpose of estimating volatility, a non-zero drift is immaterial.

Finally, it is worth noting why this control variables principle is irrelevant in the particular setting of irregular sampling recently considered by Hayashi, Jacod and Yoshida (2009). As it is manifest in the proof of their proposition 5.1., their assumption (C) allows them to compute higher order conditional moments of the ratio $\frac{W_{t_{n,i}}-W_{t_{n,i-1}}}{t_{n,i}-t_{n,i-1}}$ as if the random time intervals $(t_{n,i}-t_{n,i-1})$ were independent from the Brownian motion W. In other words, their assumption (C) precludes the kind of skewness we are taking advantage of. This assumption (C) is indeed exactly what it takes to be able to write down the likelihood function of a diffusion process irregularly sampled in time by simply plugging the random times into the diffusion transition density function. A contrario, we can make explicit by negation of their assumption (C) what we actually call endogeneity of time. Time is said endogenous because even given $F_{t_{n,i-1}}$, the variable $t_{n,i}$ is not independent of the Brownian motion W. This will be for instance the case when there is both leverage effect (instantaneous correlation between σ and W) and instantaneous causality between random times and volatility as considered by Renault and Werker (2009). We provide in the next section a list of possible models of random times, showing that some of them feature this kind of endogeneity and some others do not.

3 Various Examples and Illustration

EXAMPLE 1. (Times that are independent of the process). In the model of Mykland and Zhang (2006), the times $t_{n,i}$ are independent of the process X_t, or equivalently, nonrandom but irregularly spaced. By comparing their Proposition 1 (p. 1940) with our Theorem 1 above, it follows that $v_t \equiv 0$, and $u_t = 3\sigma_t^4 H'(t)$. Equidistant sampling is special case ($H(t) = t$). □

EXAMPLE 2. (Times generated by a fixed distortion from equidistant sampling). In Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008), times are allowed to be unequally spaced if they follow $t_{n,i} = F(i/n)$, where F is allowed to be a smooth random process which *does not depend on* n (Section 5.3, p. 1505-1507). This induces some measure of endogeneity, but not enough to avoid $v_t \equiv 0$. □

EXAMPLE 3. (Times generated by flat price trading). In the model recently proposed by Phillips and Yu (2008), the microstructure noise completely offsets the effect of price movement over the subinterval in which flat price occurs. In other words, the efficient price may be exactly observed from time to time but only at random dates defined as:

$$t_{n,i} - t_{n,i-1} = \frac{D_i}{n}$$

where (D_i) is a strictly stationary and ergodic sequence of nonnegative random variables with finite variance. These variables are allowed to depend only on past observed prices. In other words, assumption (C) of Hayashi, Jacod and Yoshida (2009) is fulfilled and thus endogeneity of time is still not enough to avoid $v_t = 0$. By a slight extension of Mykland and Zhang (2006), Phillips and Yu (2008) actually show directly that we are back to the result of Example 1. □

EXAMPLE 4. (Times generated by hitting a barrier). For simplicity, take $\mu_t \equiv 0$ and $\sigma_t \equiv 1$. The times $t_{n,i}$ are defined recursively: $t_{n,0} = 0$, and $t_{n,i+1}$ is the first time

$t \geq t_{n,i}$ so that $X_t - X_{t_{n,i}} =$ either $n^{-1/2}a$ or $-n^{-1/2}b$, where $a, b > 0$. Let N be the number of $t_{n,i} < T$, so that $t_{n,N} < T \leq t_{n,N+1}$. Redefine $t_{n,N+1} = T$.

In other words,

$$X_{t_{n,i+1}} - X_{t_{n,i}} = n^{-1/2} Z_{i+1} \text{ for } t_{n,i+1} < T, \tag{8}$$

where $Z_1, Z_2, ...$ are i.i.d. with mean zero and point mass as a and $-b$ (so $P(Z = a) = b/(a+b)$).

By standard renewal arguments $N/n \xrightarrow{p} T/(ab)$, and so the conditions of Theorem 1 are satisfied, with $v_t \equiv E(Z^3)/(ab)^{3/2}$ and $u_t \equiv E(Z^4)/(ab)^2$. We note that v_t is nonzero except when $a = b$. □

Example 5. (General return distributions). From Appendix 1 of Hall and Heyde (1980), the distribution of a general random variable (with mean zero) can be generated by the same device as in the previous example, by letting the barrier itself be random. (In mathematical terms, this is called embedding in Brownian motion.) In this more general setting, equation (8) remains valid, and the Z_i are i.i.d. with *any* mean zero distribution. If we take $E(Z^4) < \infty$, the conditions for Theorem 1 remain satisfied, and it is still the case that $v_t \equiv E(Z^3)/(E(Z^2))^{3/2}$ and $u_t \equiv E(Z^4)/(E(Z^2))^2$. □

Example 6. (Connection to the structural autoregressive conditional duration model). The paper by Renault, Van der Heijden, and Werker (2009) generalizes the hitting time technique of Example 4 above to construct autoregressive conditional duration models. It rests upon a dynamic version of Abbring (2007)'s mixed hitting time model. The observation times are defined recursively as:

$$t_{i+1} = \inf\{t > t_i : |Z_t - Z_{t_i}| > \varphi_{t_i} M_i\}$$

where Z is a Brownian motion with drift μ_Z and, for identification purpose, unit

variance. The important difference with Example 4 is that hitting barriers are now defined through a latent Brownian motion Z with drift μ_Z, which may be only partially (or not) correlated with the Brownian motion W defining price dynamics. The double-boundary setting is more convenient than a single-boundary one as it ensures that durations have finite expectations. Note that the kind of asymmetry which matters for us, namely the asymmetric barriers that yields $v_t \neq 0$, is accommodated by the non-zero correlation between the two Brownian motions Z and W, which precisely means that random times are endogenous.

More precisely, a conditional mixture feature of observed prices is produced by the mixing variables $M_i, i = 1, 2, ..n$, which are i.i.d. positive random variables with unit expectation (for reasons of identification) and M_i is independent of F_{t_i}. By contrast, the positive variable φ_{t_i} is F_{t_i}-measurable and captures observed heterogeneity in the thresholds and associated hitting times. Given both F_{t_i} and the unobserved heterogeneity M_i, the log-price process $(X_{t_i+h})_{0\leq h\leq \Delta t_{i+1}}$(with $\Delta t_{i+1} = t_{i+1} - t_i$) is specified as a Brownian motion with drift $\mu_{t_i}(M_i)$ and variance $\sigma^2_{t_i}(M_i)$. Moreover, the couple $(X_{t_i+h}, Z_{t_i+h})_{0\leq h\leq \Delta t_{i+1}}$ follows a bivariate Brownian motion with instantaneous correlation (still conditional on F_{t_i} and M_i) denoted by $\rho_{t_i}(M_i)$. It can then be shown that conditionally on F_{t_i} , M_i and Δt_{i+1}, the log-price change $X_{t_{i+1}} - X_{t_i}$ follows a mixture of two normal distributions with the same variance and respective means:

$$[\mu_{t_i}(M_i) - \rho_{t_i}(M_i)\sigma_{t_i}(M_i)\mu_Z]\Delta t_{i+1} \pm \rho_{t_i}(M_i)\sigma_{t_i}(M_i)\varphi_{t_i}M_i$$

Since a mixture of two normal distributions can feature a non-zero skewness if and only if the means in the two components are different, the announced skewness (and associated time endogeneity effect) pops up if and only if the correlation coefficient $\rho_{t_i}(M_i)$ is not zero. □

EXAMPLE 7. (Connection to uncertainty zones). Robert and Rosenbaum (2009b,a) propose a model where endogenous transaction dates are produced by the fact that the transaction prices are bound to lie on a tick grid defined by multiples $k\alpha, k \in \mathbb{N}$, of a tick size α. For a current mid-tick grid value $m_k = (k + 1/2)\alpha$, they consider an uncertainty zone $U_k = [m_k - \eta\alpha, m_k + \eta\alpha]$ for some given number $\eta, 0 < \eta < 1$. The zones U_k are called uncertainty zones since they represent bands inside of which the efficient price cannot trigger a change of the transaction price. The observation times are corresponding exit times $t_{\alpha,i}$ where for the purpose of asymptotic theory the tick size α is considered as converging to zero (analogous to $t_{n,i}$ in Example 4 with $n \to \infty$). Interestingly enough, the control variable principle of variance reduction by regression of the error on the price process works differently depending upon whether one considers the quadratic variation estimation error or the hedging error due to uncertainty zone. Since the uncertainty zones are symmetric around the mid tick values, the asymptotic theory of estimation of quadratic variation is similar to Example 4 with $a = b$. Robert and Rosenbaum (2009b) do show directly (see their Lemma 12) that there is no such thing as the bias term of our Theorem 1 because the corresponding skewness term is zero. By contrast, when it comes to hedging errors, there is some relevant asymmetry if and only if $\eta \neq 1/2$. This is due to the fact that, except if $\eta = 1/2$, when starting from one side of an uncertainty zone, the barriers to reach are asymmetric. Robert and Rosenbaum (2009a) do show directly (see their Lemma 5.8 and their Theorem 4.2.) that the (asymptotic) continuous time regression of the hedging error on the price process is non-zero if and only if $\eta \neq 1/2$. In other words, the control variable principle put forward in theorem 1 above can be fruitfully applied for variance reduction in many different contexts.

□

4 Testing for the Presence of Endogenous Times

We here present three tests for endogeneity of times. We shall see in the next section that when applied to the financial data that we consider here, all the tests reject the null hypothesis of non-endogeneity.

4.1 Test I

Under the null hypothesis (H_0) that the times t_i are independent of the process X_t, we proceed as follows.

We assume that the data are divided into J blocks of size M. For block number j, covering the time period $(t_{M(j-1)}, t_{Mj}]$, the R^2 statistic is given by

$$R_j^2 = \frac{\left(\sum_{i=M(j-1)}^{Mj-1} (\Delta X_{t_i})^3\right)^2}{\left(\sum_{i=M(j-1)}^{Mj-1} (\Delta X_{t_i})^2\right)\left(\sum_{i=M(j-1)}^{Mj-1} (\Delta X_{t_i})^4\right)}, \tag{9}$$

where $\Delta X_{t_i} = X_{t_{i+1}} - X_{t_i}$. The overall test statistic is

$$T_1 = \sum_{j=1}^{J} R_j^2 \cdot \Delta\tau_j, \tag{10}$$

where $\Delta\tau_j = t_{Mj} - t_{M(j-1)}$. Following Mykland and Zhang (2009b), the following statistic provides an asymptotically valid null-distribution:

$$T_{1,0} = \sum_{j=1}^{J} R_{j,0}^2 \cdot \Delta\tau_j, \tag{11}$$

where

$$R_{j,0}^2 = \frac{\left(\sum_{i=M(j-1)}^{Mj-1} V_i^3\right)^2}{\left(\sum_{i=M(j-1)}^{Mj-1} V_i^2\right)\left(\sum_{i=M(j-1)}^{Mj-1} V_i^4\right)}, \tag{12}$$

where $V_i = (\Delta t_i)^{1/2} \times \xi_i$, where $\Delta t_i = t_{i+1} - t_i$ and $\xi_1, ..., \xi_n$ is i.i.d. standard normal.

Note that under the alternative, by (6), on each block $(t_{M(j-1)}, t_{Mj}]$,

$$u_t \approx \frac{n \sum_{i=M(j-1)}^{Mj-1} (\Delta X_{t_i})^4}{\Delta \tau_j}, \quad \nu_t \approx \frac{\sqrt{n} \sum_{i=M(j-1)}^{Mj-1} (\Delta X_{t_i})^3}{\Delta \tau_j},$$

and

$$\sigma_t^2 \approx \frac{\sum_{i=M(j-1)}^{Mj-1} (\Delta X_{t_i})^2}{\Delta \tau_j},$$

hence, one expects, subject to regularity conditions, that, as $n \to \infty$.

$$T_1 \ \to \ \int_0^T \frac{v_t^2}{\sigma_t^2 u_t} dt. \tag{13}$$

4.2 Test II

We again assume that the data are divided into J blocks of size M. For block number j, covering the time period $(t_{M(j-1)}, t_{Mj}]$, define

$$A_j = \Delta \tau_j \cdot \frac{\left(\sqrt{n} \sum_{i=M(j-1)}^{Mj-1} (\Delta X_{t_i})^3 \right)^2}{\left(\sum_{i=M(j-1)}^{Mj-1} (\Delta X_{t_i})^2 \right)^3}, \tag{14}$$

where $\Delta \tau_j = t_{Mj} - t_{M(j-1)}$. The overall test statistic is

$$T_2 = \sum_{j=1}^{J} A_j \cdot \Delta \tau_j. \tag{15}$$

The following statistic provides an asymptotically valid null-distribution:

$$T_{2,0} = \sum_{j=1}^{J} A_{j,0} \cdot \Delta\tau_j, \tag{16}$$

where

$$A_{j,0} = \Delta\tau_j \cdot \frac{\left(\sqrt{n}\sum_{i=M(j-1)}^{Mj-1} V_i^3\right)^2}{\left(\sum_{i=M(j-1)}^{Mj-1} V_i^2\right)^3}, \tag{17}$$

where $V_i = (\Delta t_i)^{1/2} \times \xi_i$, where $\xi_1, ..., \xi_n$ are i.i.d. standard normal.

Under the alternative, subject to regularity conditions, as $n \to \infty$,

$$T_2 \ \to_P \ \int_0^T \frac{v_t^2}{\sigma_t^6}\, dt. \tag{18}$$

The main difference between Test I and Test II is that in Test II the fourth powers of returns are not used. This reduces potential effects due to outliers since higher order powers exaggerate outlier effects.

4.3 Test III

The test statistic here is

$$T_3 = \frac{\sum_{j=1}^{J}\left(\frac{2}{3}n\sum_{i=M(j-1)}^{Mj-1}(\Delta X_{t_i})^4 - \frac{4}{9}\frac{\left(\sqrt{n}\sum_{i=M(j-1)}^{Mj-1}(\Delta X_{t_i})^3\right)^2}{\left(\sum_{i=M(j-1)}^{Mj-1}(\Delta X_{t_i}^2)\right)}\right)}{2\sum_{j=1}^{J}\frac{\left(\sum_{i=M(j-1)}^{Mj-1}(\Delta X_{t_i})^2\right)^2}{(\Delta\tau_j)^2}\cdot n\sum_{i=M(j-1)}^{Mj-1}(\Delta t_i)^2}.$$

Note that the numerator is an estimator of the asymptotic variance given by Theorem 1; and the denominator is estimating $2\int_0^1 \sigma_s^4\, dH_s$, which, *under the non-endogenous hypothesis*, equals the asymptotic variance.

Replacing ΔX_{t_i} with $V_i = (\Delta t_i)^{1/2} \times \xi_i$ where $\xi_1, ..., \xi_n$ are i.i.d. standard normal provides an asymptotically valid null-distribution:

$$T_{3,0} = \frac{\sum_{j=1}^{J} \left(\frac{2}{3} n \sum_{i=M(j-1)}^{Mj-1} V_i^4 - \frac{4}{9} \frac{\left(\sqrt{n} \sum_{i=M(j-1)}^{Mj-1} V_i^3 \right)^2}{\left(\sum_{i=M(j-1)}^{Mj-1} V_i^2 \right)} \right)}{2 \sum_{j=1}^{J} \frac{\left(\sum_{i=M(j-1)}^{Mj-1} V_i^2 \right)^2}{(\Delta \tau_j)^2} \cdot n \sum_{i=M(j-1)}^{Mj-1} (\Delta t_i)^2}.$$

Under null with the conventional assumption (1),

$$T_3 \to \frac{\int_0^t \frac{2}{3} \cdot 3\sigma_s^4 \, dH_s}{2 \int_0^t \sigma_s^4 \, dH_s} = 1;$$

under alternative,

$$T_3 \to \frac{\int_0^t \left(\frac{2}{3} u_s - \frac{4}{9} \frac{v_s^2}{\sigma_s^2} \right) \, ds}{2 \int_0^t \sigma_s^4 \, dH_s}.$$

4.4 Combining Several Days

Each of the above tests can be used to test the presence of endogenous times. When the p-values are independent over days (or have approximate martingale structure), we can combine all the p-values and obtain a combined p-value using Fisher's combined test. More explicitly, if we let p_i $(i = 1, \ldots, N)$ be the p-values from day 1 to day N, then under the null

$$-2 \sum_{i=1}^{N} \log(p_i) \sim \chi^2_{2N}.$$

We can then compare $-2 \sum_{i=1}^{N} \log(p_i)$ with the χ^2_{2N} distribution and get a combined p-value

$$P_{combined} = P \left(\chi^2_{2N} > -2 \sum_{i=1}^{N} \log(p_i) \right).$$

5 Empirical Study

5.1 Data Description

We use trade data from the TAQ database. We consider several traded stocks at NYSE. Our analysis is based on subsampled local-averaged log prices. More specifically, we sample every K time stamps; for each time stamp in this sub-grid, we use the average of its preceding P observations *in the original complete price record* and treat the subsampled local-averaged log price as the log price at that time point, and we take $P < K$ so there is no overlapping. Note that the local-averaging is a modified version of the "pre-averaging" (Jacod, Li, Mykland, Podolskij, and Vetter (2009)), which can be considered as a way to reduce microstructure noise. We are using local-averaging ($P << K$) instead of pre-averaging ($P = K$) in order to retain more precise information of the observation times. We consider only the transactions within the 9 : 30 am to 4 pm window when the exchange is open. Note that we should choose K large enough so that there is almost no multiple observations sharing the same timestamp (so that the Δt_i's are reasonably precise).

Mathematically, suppose the raw data is $X^o_{v_\ell}$, $\ell = 1, \cdots, L$. Our analysis will be based on X_{t_i} with

$$t_i = v_{(i-1)K+P}, \quad i = 1, \cdots n = \lfloor \frac{L}{K} \rfloor,$$

and

$$X_{t_i} = \frac{1}{P} \sum_{j=1}^{P} X^o_{v_{(i-1)K+j}}.$$

We conduct the tests mentioned in Section 5.1-5.3. For Tests I and II, onesided

p-values make sense, because asymptotically, the test statistics converge to zero under null, but to positive numbers under alternative. For Test III, we know that the limit under null is 1, so we use two-sided p-values. More explicitly, for either Test I or Test II, for each day i, we simulate n null statistics $T_{i,j}$ $(j = 1, \ldots, n)$ ($n = 1000$ in the following study) and the estimated one-sided p-value is

$$\hat{p}_i = \max\left(\frac{\sum_j \mathbf{1}_{\{T_{i,j} > T_i\}}}{n}, \frac{1}{n}\right);$$

for Test III,

$$\hat{p}_i = \max\left(\frac{\sum_j \mathbf{1}_{\{|1-T_{i,j}| > |1-T_i|\}}}{n}, \frac{1}{n}\right).$$

We check the acf plots of the p-values for the independence, and then use Fisher's combined test (see Section 4.4) to find the combined p-values.

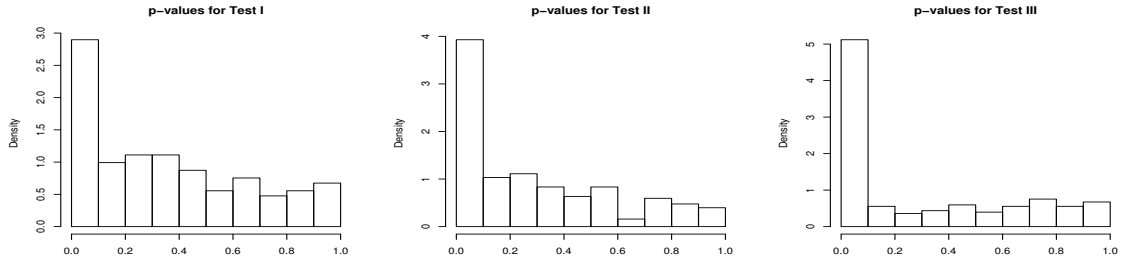
5.2 Test Results

We here study the behavior of our test statistics for four stocks: SKS, DDS, MAT and IBM. We show the distribution of daily p-values for one year (SKS and DDS) or 3 months (MAT and IBM), along with a combined p-value (Section 5.4). It is clear from the results that the null hypothesis of non-endogeneity is rejected for all the stocks and all the statistics when aggregated over the total time period. The result may vary over individual days, either due to statistical variability or to the varying dynamics. Though not strictly needed, we also provide autocorrelation function (ACF) plot of the p-values to show that they are uncorrelated across days.

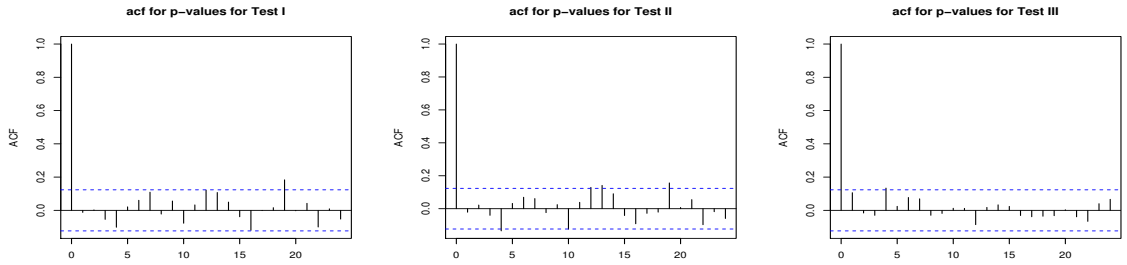
5.2.1 SKS

SKS 2005 one year data. Parameters used: local-averaging scheme $P = 3$, subsample scheme $K = 8$ and number of blocks $J = 3$ (block size $M \approx 50$).

Histograms of the p-values:



Check of independence between the p-values:



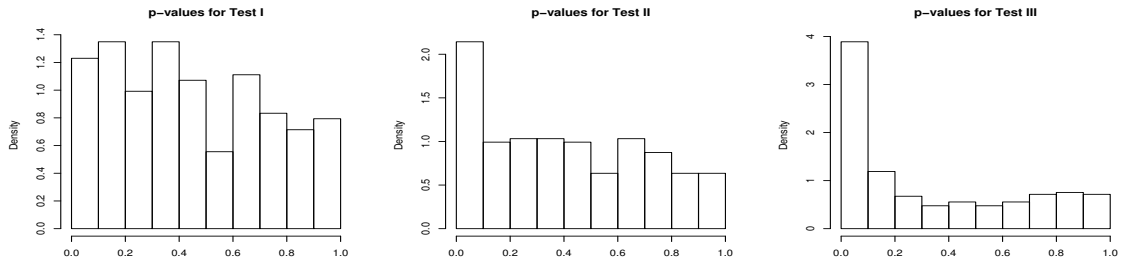
Combined p-values (total time period, see Section 4.4):

Tests	I	II	III
p-values	0	0	0

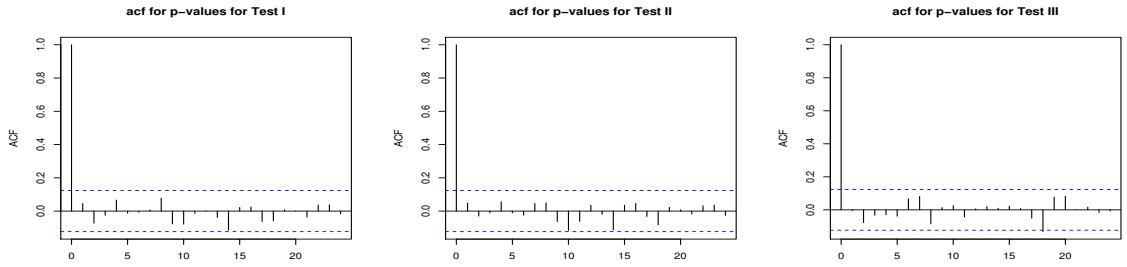
5.2.2 DDS

DDS 2005 one year data. Parameters used: local-averaging scheme $P = 3$, subsample scheme $K = 8$ and number of blocks $J = 3$ (block size $M \approx 50$).

Histograms of the p-values:



Check of independence between the p-values:



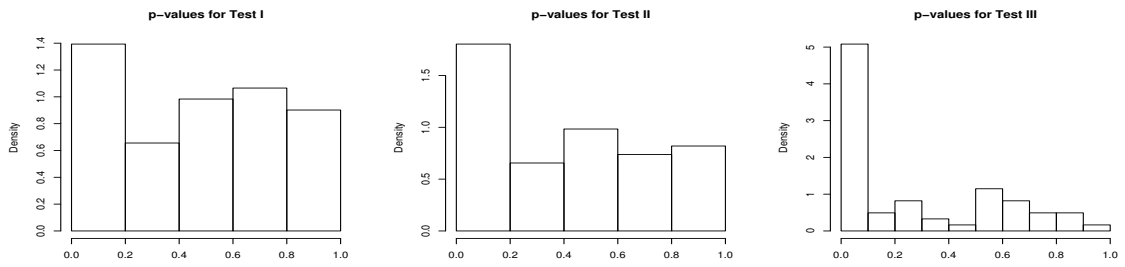
Combined p-values (total time period):

Tests	I	II	III
p-values	0.0001	5.55e-15	0

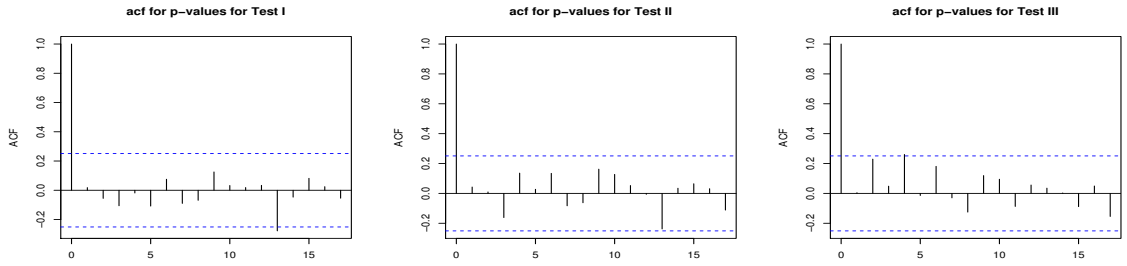
5.2.3 MAT

MAT 2005 Jan-Mar three months' data. Parameters used: local-averaging scheme $P = 3$, subsample scheme $K = 8$ and number of blocks $J = 5$ (block size $M \approx 50$).

Histograms of the p-values:



Check of independence between the p-values:



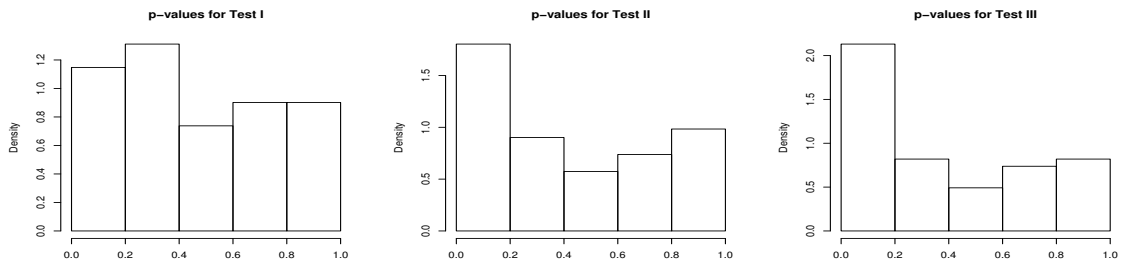
Combined P-values (total time period):

Tests	I	II	III
p-values	0.124	4.35e-07	0

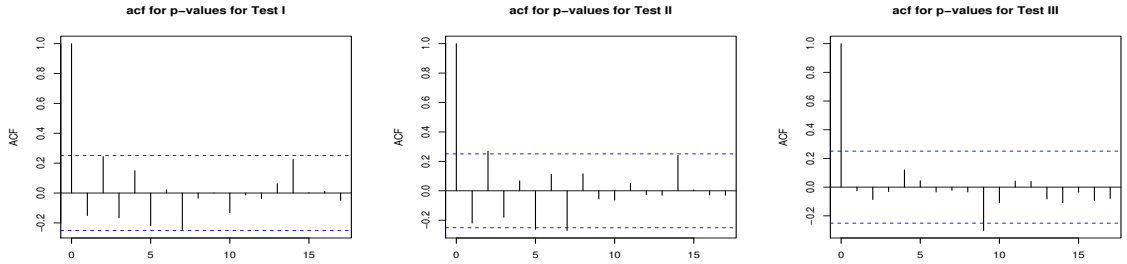
5.2.4 IBM

IBM 2005 Jan-Mar three months' data. Parameters used: local-averaging scheme $P = 5$, subsample scheme $K = 20$ and number of blocks $J = 5$ (block size $M \approx 65$).

Histograms of the p-values:



Check of independence between the p-values:



Combined p-values (total time period):

Test	I	II	III
p-values	0.093	3.82e-05	1.15e-10

6 Simulation Study

We took the same setting as in Example 3 in Section 3 with $\mu = 0$, $\sigma = 0.02$, $a = 0.04$, $b = 0.01$ and $n = 3600$. If we think of the simulated process as a log price process, then the stopping rule makes that there is a transaction each time when there is an increase of 0.067% or a decrease of 0.017%. The actual number of daily trades is about 3000.

We examine three confidence intervals based on three different methods. Confidence intervals of 22 days are plotted in the upper panel of the Figure 1. Confidence intervals of 1000 days are plotted in the lower panel of Figure 1.

- Confidence intervals CI_H (green dashed lines). These are built out of the naive method ignoring the dependency between the observation times and the process, using the CLT based on the quadratic variation of times:

$$\sqrt{n}\left(RV_T - \int_0^T \sigma_t^2 dt\right) \to_{\mathcal{L}-Stably} \int_0^T \sqrt{2\sigma_t^4 H'(t)}\, dB_t,$$

  where $B_t \perp\!\!\!\perp W_t$, and H_t is defined by (1) (it is usually assumed to be differentiable).

- Confidence intervals CI_X (blue dotted lines). These are built by still ignoring the dependency between the observation time and the process, but using the CLT based on the realized quarticity which is equivalent to the above CLT if there were no endogeneity:

$$\sqrt{n}\left(RV_T - \int_0^T \sigma_t^2 dt\right) \to_{\mathcal{L}-Stably} \int_0^T \sqrt{\frac{2}{3}u_s}\, dB_t,$$

  where $B_t \perp\!\!\!\perp W_t$.

- Confidence intervals CI_C (red solid lines). These are based on Theorem 1, by first estimating the asymptotic bias, and then correcting for it from the Realized Volatility. The variance is corrected accordingly.

In estimating the processes σ_s, H_s, u_s and v_s, we use the block method as in Section 4.1. The number of blocks is chosen to be $J = 3$, which corresponds to a block size of $M \approx 1000$.

Remark 1. The choice of $J = 3$ is not optimal. We are acting as if we did not know how the data were generated; otherwise we would choose $J = 1$, because in this setting, the processes σ_s, u_s and v_s are all constant over the whole time period, hence putting all data in one block gives the smallest errors. In practice, one can use this idea as a

guidance to pick J in a bootstrap manner; we shall discuss this in a subsequent paper.

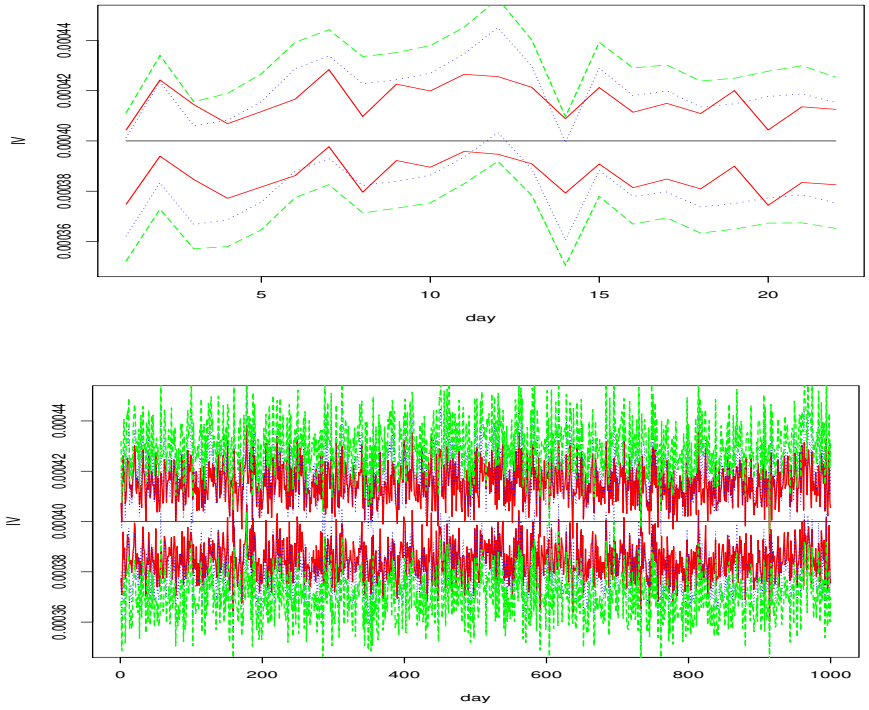


Figure 1. Confidence intervals computed based on the three methods explained above (Green dashed: CI_H; Blue dotted: CI_X; Red solid: CI_C). Upper panel: for 22 days; lower panel: for 1000 days.

Summary statistics (based on simulation of 1000 days):

	Average width	RMSE	Coverage Frequency	% Reduced width compared with CI_H	% Reduced RMSE compared with CI_H
CI_H	6.161e-05	1.005e-05	99.6%	–	–
CI_X	4.020e-05	1.005e-05	95.8%	34.7%	0
CI_C	3.013e-05	7.460e-06	95.5%	51.1%	25.8%

The RMSE in the table above stands for the root mean of the squared distance between the centers of the confidence intervals and the true σ^2.

From the plots and the summary statistics we have the following observations.

1. Width of the confidence intervals: We see that CI_X is much narrower than CI_H. This reflects the fact that in the endogenous case the asymptotic variance $\lim_n \frac{2}{3}n[X,X,X,X]_t$ may be substantially different from $\int_0^t 2\sigma_s^4 dH_s$ which is the asymptotic variance one would get if the endogeneity is overlooked. Furthermore, the correct confidence interval CI_C is even narrower than CI_X.

2. Bias correction: When the blue confidence intervals tend to be too extreme and not covering the true value, our bias correction may correct it back especially when the extremeness of the blue confidence interval was due to the dependency of the time and process rather than pure randomness.

3. Coverage frequency: We see from the summary statistics that the confidence intervals CI_C have coverage frequency of 95.5%, and in the mean while being narrower than the confidence intervals based on the other two methods. This coverage frequency is close to what is being expected (95%), and is similar to that achieved by the CI_X, which are wider. Despite the bias, the CI_H have bigger coverage frequency which is mainly due to the (wrongly estimated) bigger width.

7 Conclusion

We have established a central limit theorem for general dependent times. We also show that the endogeneity can exist in financial data, using tests based on our theory. It remains an open question how to estimate the size of the effect, and this is deferred to later work.

REFERENCES

Abbring, J. H. (2007), "Mixed hitting-time models," Tinbergen Institute Discussion Paper (TI2007-057/13).

Aït-Sahalia, Y. and Mykland, P. A. (2003), "The Effects of Random and Discrete Sampling When Estimating Continuous-Time Diffusions," *Econometrica*, 71, 483–549.

Aldous, D. J. and Eagleson, G. K. (1978), "On Mixing and Stability of Limit Theorems," *Annals of Probability*, 6, 325–331.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008), "Designing realized kernels to measure ex-post variation of equity prices in the presence of noise," *Econometrica*, 76, 1481–1536.

Barndorff-Nielsen, O. E. and Shephard, N. (2001), "Non-Gaussian Ornstein-Uhlenbeck-Based Models And Some Of Their Uses In Financial Economics," *Journal of the Royal Statistical Society, B*, 63, 167–241.

— (2002), "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society, B*, 64, 253–280.

Dellacherie, C. and Meyer, P. (1982), *Probabilities and Potantial B*, Amsterdam: North-Holland.

Duffie, D. and Glynn, P. (2004), "Estimation of Continuous-Time Markov Processes Sampled at Random Times," *Econometrica*, 72, 1773–1808.

Engle, R. F. (2000), "The Econometrics of Ultra-High Frequency Data," *Econometrica*, 68, 1–22.

Fukasawa, M. (2009), "Realized volatility with stochastic sampling," (discussion paper).

Grammig, J. and Wellner, M. (2002), "Modeling the interdependence of volatility and inter-transaction duration processes," *Journal of Econometrics*, 0, 0–0.

Hall, P. and Heyde, C. C. (1980), *Martingale Limit Theory and Its Application*, Boston: Academic Press.

Hayashi, T., Jacod, J., and Yoshida, N. (2008), "Irregular sampling and cenral limit theorems for power variations: the continuous case," *working paper*.

Jacod, J. (1994), "Limit of Random Measures Associated with the Increments of a Brownian Semimartingale," Tech. rep., Université de Paris VI.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009), "Microstructure Noise in the Continuous Case: The Pre-Averaging Approach," *Stochastic Processes and Their Applications*, 119, 2249–2276.

Jacod, J. and Protter, P. (1998), "Asymptotic Error Distributions for the Euler Method for Stochastic Differential Equations," *Annals of Probability*, 26, 267–307.

Meddahi, N., Renault, E., and Werker, B. (2006), "GARCH and Irregularly Spaced Data," *Economics Letters*, 90, 200–204.

Mykland, P. A. and Zhang, L. (2006), "ANOVA for Diffusions and Itô Processes," *Annals of Statistics*, 34, 1931–1963.

— (2009a), "The Econometrics of High Frequency Data," (to appear in *Statistical Methods for Stochastic Differential Equations*, M. Kessler, A. Lindner, and M. Sørensen, eds., Chapman and Hall/CRC Press).

— (2009b), "Inference for continuous semimartingales observed at high frequency," *Econometrica*, 77, 1403–1455.

Phillips, P. C. B. and Yu, J. (2007), "Information Loss in Volatility Measurement with Flat Price Trading," *working paper*.

Protter, P. (2004), *Stochastic Integration and Differential Equations: A New Approach*, New York: Springer-Verlag, 2nd ed.

Renault, E., Van der Heijden, T., and Werker, B. J. (2009), "A structural autoregressive conditional duration model," (in preparation).

Renault, E. and Werker, B. J. (2009), "Causality effects in return volatility measures with random times," *Journal of Econometrics* (forthcoming).

Rényi, A. (1963), "On Stable Sequences of Events," *Sankyā Series A*, 25, 293–302.

Robert, C. Y. and Rosenbaum, M. (2009a), "On the Microstructural Hedging Error," W.P. CMAP, Ecole Polytechnique, Paris.

— (2009b), "Volatility and Covariation Estimation when Microstructure Noise and Trading Times are Endogenous," W.P. CMAP, Ecole Polytechnique, Paris.

Rootzén, H. (1980), "Limit Distributions for the Error in Approximations of Stochastic Integrals," *Annals of Probability*, 8, 241–251.

Zhang, L. (2001), "From Martingales to ANOVA: Implied and Realized Volatility," Ph.D. thesis, The University of Chicago, Department of Statistics.

A Appendix: Proof of Theorem 1

Because we shall prove stable convergence, and because of the local boundedness, we can without loss of generality assume that σ_t and μ_t are bounded by a nonrandom constant. One can further suppress μ as in Section 2.2 (p. 1407-1409) of Mykland and Zhang (2009b), and act as if X is a local martingale.

Define the interpolated and rescaled error process by

$$dM_t = 2n^{1/2}(X_t - X_{t_*})dX_t \ , \quad M_0 = 0.$$

where t_* is the largest time t_i smaller than or equal to t. From (6), it follows as in the proof of Proposition 2 (p. 1952) of Mykland and Zhang (2006) that $\langle M, M\rangle_t \xrightarrow{p} \frac{2}{3}\int_0^t u_s ds$ for all t (the proof does not depend on times being nonrandom). The remainder term in equation (6.3) of that paper vanishes at the relevant order because of our condition (5).

Specifically, this works as follows. With the same interpolation of $[X, X, X, X]_t$, and using the first part of equation (6.3) in Mykland and Zhang (2006), we obtain

$$nd[X, X, X, X]_t = \frac{3}{2}d\langle M, M\rangle_t + 4n(X_t - X_{t_*})^3 dX_t. \tag{A.1}$$

Without loss of generality, we can assume that σ_t is bounded by, say, σ_+ (see Section 4.5 of Mykland and Zhang (2009a)). By the Burkholder-Davis-Gundy inequality (see Section 3 of Ch. VII of Dellacherie and Meyer (1982), or p. 193 and 222 in Protter

(2004)), the expected quadratic variation of the second term satisfies

$$\begin{aligned} E\langle \int_0^{\cdot} 4n(X_t - X_{t_*})^3 dX_t, \int_0^{\cdot} 4n(X_t - X_{t_*})^3 dX_t \rangle_T &= 16n^2 E \int_0^T (X_t - X_{t_*})^6 d\langle X, X\rangle_t \\ &\leq c16n^2\sigma_+^8 E \int_0^T (t - t_*)^3 dt \\ &\leq c16Tn^{-3\epsilon}\sigma_+^8 \\ &\xrightarrow{p} 0 \ , \end{aligned}$$

where c is a universal constant, and where the second-to-last transition is by assumption (5). Having eliminated the second term in (A.1), it follows that $\langle M, M\rangle_t \xrightarrow{p} \frac{2}{3}\int_0^t u_s ds$ for all t.

Similarly, (7) yields that $\langle X, M\rangle_t \xrightarrow{p} \frac{2}{3}\int_0^t v_s ds$, again for all t. The analogous equation to (A.1) follows from Itô's formula since

$$\begin{aligned} d(X_t - X_{t_*})(M_t - M_{t_*}) &= d\langle X, M\rangle_t + \text{ martingale term} \\ &= 2n^{1/2}(X_t - X_{t_*}) d\langle X, X\rangle_t + \text{ martingale term} \\ &= \frac{2}{3} n^{1/2} d(X_t - X_{t_*})^3 + \text{ martingale term.} \end{aligned}$$

The martingale term is negligible again by (5).

The overall result now follows from the limit results in either Theorem B.4 (p. 65-67) of Zhang (2001), or Theorem 6 of Mykland and Zhang (2009a).

□