Realized Volatility When Sampling Times are Possibly Endogenous *

Yingying Li
Hong Kong University of Science and Technology

Per A. Mykland
University of Chicago

Eric Renault
Brown University

Lan Zhang
University of Illinois at Chicago

Xinghua Zheng
Hong Kong University of Science and Technology

This version: April 24, 2013.

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

*We are grateful to Andrew Patton and Neil Shephard, and the participants of the Stevanovich Center - CREATES 2009 conference and the Fourth Annual SoFiE Conference for their comments and suggestions. Financial support from the Research Grants Council of Hong Kong under grants GRF 602710 and GRF 606811 (Li and Zheng), and the National Science Foundation under grants DMS 06-04758, SES 06-31605, and SES 11-24526 (Mykland and Zhang) 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.

also establish a central limit theorem for the tricity under the hypothesis that there is no endogeneity, based on which we propose a test and 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 Itô process $dX_t = \mu_t dt + \sigma_t dW_t$ (with W_t Wiener process) observed at times $t_{n,i},\ i = 0, 1, \ldots, N$, in the time interval $[0, 1]$, where n characterizes the observation frequency, and $N = N_n$, which may be random, stands for the actual number of observations before time 1. One can think of n to be the expected number of observations per period or something similar. In the non-endogenous case, i.e., when observation times are independent of the price process, without loss of generality, we can and we will take $n = N$. More generally, if one assumes that N/n has a (possibly random) probability limit F when n goes to infinity,[1] we will be able to establish a feasible asymptotic theory in terms of N, see Theorem 1. The process X_t must be understood as a log-price so that $\Delta X_{t_{n,i}} := X_{t_{n,i}} - X_{t_{n,i-1}}$ 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 $\max_i \Delta t_{n,i} := \max_i(t_{n,i} - t_{n,i-1})$ goes to zero in probability, then the realized

[1] This is a natural assumption in view of our examples, and also from renewal theory type considerations; see, for example, Ross (1996).

variance $[X,X]_t = \sum_{t_{n,i} \le t} \Delta X^2_{t_{n,i}}$ is a consistent estimator of the quadratic variation $\langle X,X\rangle_t = \int_0^t \sigma_s^2 ds$.

Second, if the times $t_{n,i}$'s are independent of the X process, and 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,

$$\lim_{n\to\infty} N \sum_{t_{n,i}\le t} \Delta t_{n,i}{}^2 = H_t, \tag{1}$$

where H_t is an adapted process,[2] 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 $2\int_0^t \sigma_s^4 \; dH_s$. It is also known that in this case $\int_0^t \sigma_s^4 \; dH_s$ is consistently estimated by

$$\frac{N}{3}[X,X,X,X]_t = \frac{N}{3} \sum_{t_{n,i}\le t} \Delta X^4_{t_{n,i}}, \tag{2}$$

(Barndorff-Nielsen and Shephard (2002))[3]. In the equidistant case, i.e., when $t_{n,i} = i/n$, (1) holds with $H_t = t$. The main intuition is that the Itô process X is locally conditionally Gaussian and thus features a kurtosis coefficient equal to 3.

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 (2011) and Phillips and Yu (2007), but also when there is no asymptotic bias.

[2]Here and in the sequel, when times are assumed to be independent of the process X_t, one may assume that the times are measurable with respect to the time zero sigma-field.

[3] For readability, we shall refer to both (2) and the unscaled version $[X,X,X,X]_t$ as quarticity. The same will apply to the tricity, introduced below in (3).

A striking feature of these results is that the *tricity*

$$[X, X, X]_t = \sum_{t_{n,i} \leq t} \Delta X^3_{t_{n,i}} \tag{3}$$

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 $t_{n,i} = i/n$, we have $N = n$ and

$$N^{1/2}[X, X, X]_t = n^{1/2}\sigma^3 \sum_{t_{n,i} \leq t} (W_{t_{n,i}} - W_{t_{n,i-1}})^3 =_{\mathcal{L}} \sigma^3 \frac{1}{n} \sum_{i=1}^{[nt]} Z_i^3,$$

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

$$\lim_{n \to \infty} N^{1/2}[X, X, X]_t = 0. \tag{4}$$

By a standard predictability argument, the property (4) 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 $N^{1/2}[X, X, X]_t$ (see, e.g., Mykland and Zhang (2009), Example 3, p. 1414-16). 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 (i.e., independent of n) random time change (see e.g. Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008)) will not destroy the result (4). The same applies to the results of Hayashi, Jacod, and Yoshida (2011) and Phillips and Yu (2007). A maintained assumption in their setting (see use of assumption (C) in Hayashi, Jacod, and Yoshida (2011)) is that higher order conditional moments of ratios $\frac{X_{t_{n,i}}-X_{t_{n,i-1}}}{t_{n,i}-t_{n,i-1}}$ can be computed as if the random time intervals $(t_{n,i}-t_{n,i-1})$ were independent from the Brownian motion W. Then, we have locally a zero conditional skewness and a conditional kurtosis equal to 3, implying that again the tricity $N^{1/2}[X,X,X]_t$ has a zero-limit and the limit of quarticity $\frac{N}{3}[X,X,X,X]_t$ coincides with $\int_0^t \sigma_s^4 \, dH_s$. Note that this assumption (C) is indeed exactly what it takes to be able to write down a likelihood function for a diffusion process irregularly sampled in time by simply plugging the random times into the diffusion transition density function. A contrario, we call endogeneity of time a situation in which randomness in observation times does matter because it implies a non-zero limit for tricity $N^{1/2}[X,X,X]_t$ and/or a limit for quarticity $\frac{N}{3}[X,X,X,X]_t$ that does not coincide anymore with $\int_0^t \sigma_s^4 \, dH_s$.

The focus of interest of this paper are the relevant changes to make in the limit distribution of the normalized estimation error $N^{1/2}\left([X,X]_t - \langle X,X\rangle_t\right)$ to properly take into account the aforementioned effects of time endogeneity. According to our main theoretical result, the change is threefold:

First, when the process X has a non-zero drift μ, the non-zero limit for $N^{1/2}[X,X,X]_t$ will imply a non-zero mean for the asymptotic distribution of the normalized error $N^{1/2}\left([X,X]_t - \langle X,X\rangle_t\right)$.

Second, even when the process X has no drift, the asymptotic distribution of the normalized error entails a term that we can call "bias" since it can be consistently estimated for higher order improvements of our estimator of quadratic variation. Up to the drift-induced bias, the total asymptotic mean squared error is still equal to the limit of $\frac{2N}{3}[X,X,X,X]_t$ but is now decomposed as the sum of a squared bias term and a residual variance. Estimating and eventually subtracting the estimated bias term will allow to get an estimator more accurate than the usual realized variance since it reaches the efficiency bound given by the aforementioned residual variance. We then take advantage of the endogeneity bias for an improved estimator of quadratic variation if and only if the limit of the tricity $N^{1/2}[X,X,X]_t$ is not zero.

Third, it must be kept in mind that the total mean squared error, while still given (in the no-drift case) by two thirds of the limit of the normalized quarticity $N[X,X,X,X]_t$, is no longer necessarily equal to $2\int_0^t \sigma_s^4 \, dH_s$.

The bottom-line is that 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 (2011)) 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, in a very nice paper by Fukasawa (2010b). (Fukasawa (2010a) considered the special case when the observation times are passage times like in Examples 4 and 6 below, and proved a CLT for realized volatility.) A major difference between the main results in Fukasawa (2010b) and our CLT (Theorem 1 below) lies in how the main assumptions are imposed. In Fukasawa (2010b), the main assumptions (Condition 3.5 therein) are put on the increments of the martingale components of the price process, and these need to be valid locally (for "spot" conditional moments). In our CLT, the main assumptions are put on the (observable) observation time process and on integrated moments of the log-returns. In terms of estimating the bias and variance in the CLT, Fukasawa (2010b) discusses the constant skewness case. We illustrate how one can estimate the bias and variance in the general case using the blocking method as proposed by Mykland and Zhang (2009), and use them to further build confidence intervals, see Sections 4.1 and 5 for more details.

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 $\lim_{n\to\infty} N^{1/2}[X,X,X]_1$ is not zero. $\lim_{n\to\infty} N^{1/2}[X,X,X]_1$ can actually be interpreted in terms of a measure of covariance between process and time, see Remark 3.

As extensively discussed by Renault and Werker (2011), 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. In Section 4, we devise a CLT for the tricity, which we use to construct a test for the null hypothesis of non-endogeneity of the times. A simulation study is carried out in Section 5 and empirical results in Section 6. The proof of the main theorem is in the Appendix.

2 Main Result

We use the standard Itô process model for the log-price $X = (X_t)$:

$$dX_t = \mu_t \, dt + \sigma_t \, dW_t, \tag{5}$$

where W_t is a Wiener process, μ and σ take values in $\mathbb{D}[0,1]$ (the space of real-valued functions on $[0,1]$ that are right continuous and have left limits (càdlàg)), and furthermore σ_t is strictly positive. The target of inference is

$$\langle X, X\rangle_t = \int_0^t \sigma_s^2 \, ds. \tag{6}$$

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 $\mathcal{F}_1$-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 $\mathcal{F}_1$ so that for all $A \in \mathcal{F}_1$ and for all bounded continuous function g, $E(\mathbf{1}_A g(Z_n)) \to E(\mathbf{1}_A g(Z))$ as $n \to \infty$, where $\mathbf{1}_A$ is the indicator function of A.*

Since we consider convergence of processes in $\mathbb{D}[0,1]$, continuity of test functions g is normally defined with respect to the Skorokhod topology on this space. However, since all our limits are in $\mathbb{C}[0,1]$, we can also take continuity to be given by the sup-norm, cf. Chapter VI of Jacod and Shiryaev (2003). 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. *Suppose that μ_t and σ_t^2 are adapted to a filtration $(\mathcal{F}_t)$, integrable, locally bounded, and that $\inf_{t\in(0,1]}\sigma_t > 0$ almost surely. Also assume*[4] *that for some $\varepsilon > 0$,*

$$\max_i \Delta t_{n,i} = o_p(n^{-(\frac{2}{3}+\varepsilon)}). \tag{7}$$

Further assume that (for all t)

$$n[X,X,X,X]_t \xrightarrow{p} \int_0^t \widetilde{u}_s \sigma_s^4 \, ds \quad \text{and}$$
$$n^{1/2}[X,X,X]_t \xrightarrow{p} \int_0^t \widetilde{v}_s \sigma_s^3 \, ds,$$

where $[X,X,X,X]_t$ and $[X,X,X]_t$ are defined in (2) and (3) respectively, and $\widetilde{u}_t\sigma_t^4$, $\widetilde{v}_t\sigma_t^3$ and $\widetilde{v}_t^2\sigma_t^4$ are integrable. Finally, assume that the filtration $(\mathcal{F}_t)$ is generated by finitely many Brownian motions. Then, stably in law as $n \to \infty$:

$$n^{1/2}\left([X,X]_t - \langle X,X\rangle_t\right) \to \underbrace{\frac{2}{3}\int_0^t \widetilde{v}_s\sigma_s dX_s}_{\text{asymptotic bias}} + \int_0^t \sqrt{\frac{2}{3}\widetilde{u}_s - \left(\frac{2}{3}\widetilde{v}_s\right)^2}\,\sigma_s^2 \, dB_s \tag{8}$$

where B_t is a Brownian-motion independent of the underlying σ-field. In particular, if

[4]Assumption (7) weakens the condition that $\max_i \Delta t_{n,i} = O_p(n^{-1})$, which has been common in the literature. For example, our assumption permits times to be generated as a Poisson type point process while still including the earlier more restrictive assumption.

$N/n \xrightarrow{p} F$ for some (positive) random variable F, then

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

where $v_s = \sqrt{F}\widetilde{v}_s$ and $u_s = F\widetilde{u}_s$ are such that

$$N[X,X,X,X]_t \xrightarrow{p} \int_0^t u_s \sigma_s^4\, ds \quad and \quad (10)$$

$$N^{1/2}[X,X,X]_t \xrightarrow{p} \int_0^t v_s \sigma_s^3\, ds. \quad (11)$$

Observe that even if $F \notin \mathcal{F}_t$, the (stochastic) integrals on the RHS of (9) still make sense because they are multiples of the random variable $\sqrt{F}$ and the (stochastic) integrals on the RHS of (8).

Remark 1. In practice, choosing n to be, say, 1000 or 2000, does not make a difference in applying either (8) or (9), because a multiple of n leads to multiples of finite sample estimates of $(\widetilde{u}_s, \widetilde{v}_s)$ or (u_s, v_s), and the multiples cancel. □

This new result does not change the integrated variance (quadratic variation) of the limiting process, which remains $(2/3)\int_0^t u_s\sigma_s^4 ds$. It does, however, reapportion some of this quadratic variation from asymptotic variance to asymptotic mean.

Remark 2. If the coefficients μ_t, σ_t, u_t and v_t are independent of the driving Brownian motion W, then we can use an alternative form of convergence, which is conditional on the coefficients. In this case, the limit reduces to a normal distribution with mean $\frac{2}{3}\int_0^t v_s\sigma_s\mu_s ds$ and variance $(2/3)\int_0^t u_s\sigma_s^4 ds$. Hence, if $\mu_t \equiv 0$, we are back to the standard result. However, if μ_t is nonzero, this limit is infeasible. We are thus better off using the stable convergence in Theorem 1. □

Remark 3. As discussed in the introduction, there is a tight connection between endogeneity of observation times (non-zero covariance between price process and time) and a non-zero probability limit of tricity. To see this, let us assume that σ_t^2 itself is an Itô process. The connection is then that:

$$n^{1/2} \sum_{t_{n,i} \leq t} \Delta t_{n,i} \Delta X_{t_{n,i}} \xrightarrow{p} \frac{\int_0^t \widetilde{v}_s \sigma_s ds}{3}, \text{ hence } N^{1/2} \sum_{t_{n,i} \leq t} \Delta t_{n,i} \Delta X_{t_{n,i}} \xrightarrow{p} \frac{\int_0^t v_s \sigma_s ds}{3}. \quad (12)$$

This result is shown in Appendix A.2. □

Remark 4. (Jumps, Bi-and Multipower, Strong Representation.) In the case of finitely many jumps, these can be removed as in Mancini (2001) and Lee and Mykland (2008). Since the jumps are removed for large n, our results go through unchanged. The infinitely many jumps case is complex and beyond the scope of this paper.

An alternative approach to jumps makes use of bi- and multipower variation (Barndorff-Nielsen and Shephard (2004), Barndorff-Nielsen et al. (2006), Barndorff-Nielsen et al. (2006)), which can directly estimate volatility in a way that is robust to jumps. An efficient block based theory is provided in Mykland, Shephard and Sheppard (2012). The latter paper shows a strong approximation between multipower and realized volatility, see Theorem 4 (p. 12) of the paper. Under suitable regularity conditions, this strong representation will generalize to our setting of endogenous times.[5] One can then proceed as in the transition from Theorem 4 to Theorem 5 in Mykland, Shephard and Sheppard (2012), to obtain that the blocked multipower variation has the same asymptotic distribution as provided in our Theorem 1. This result is thus a corollary

[5]The volatility of volatility term in Theorem 4 of Mykland, Shephard and Sheppard (2012) must be replaced by an $O_p(n^{\frac{3}{2}-\beta} \max \Delta t_{n,i})$ term, but for β close enough to 1, the merger with our current Theorem 1 goes through.

to the current theorem and to a slight extension of the other paper. □

Remark 5. An interesting open question concerns the extension to multivariate processes. If the observation times are the same for all the dimensions, this extension is straightforward. However, a more realistic set of assumptions would involve endogenous and also asynchronous times. As is known from the literature on exogenous times, this is a complicated problem. See, for example, Hayashi and Yoshida (2005), and other work by the same authors. See also Christensen, Podolskij and Vetter (2011) and Zhang (2011) for the case with microstructure noise. □

Remark 6. The paper does not study the case with microstructure noise, and in implementation we have relied on sparse sampling, which is conventional in large parts of the literature. The effect of microstructure noise on realized volatility itself is similar to that of Section 2 of Zhang, Mykland and Aït-Sahalia (2005) (ZMA), with the adjustment that the right hand side of equation (25) in ZMA has to be modified to have the distribution of the current Theorem 1. Proposition 1 in ZMA is modified similarly.

Approaches to dealing with microstructure noise through modification of realized volatility include (in the exogenous-time case) Zhang, Mykland and Aït-Sahalia (2005), Zhang (2006), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008), Jacod, Li, Mykland, Podolskij, and Vetter (2009), and Xiu (2010). We do not know how these procedures work under endogenous time. The work of Robert and Rosenbaum (2010, 2012) provides a different angle on microstructure noise, and is discussed in Example 7 below. □

Theorem 1 suggests an improvement of the classical estimator $[X, X]_t$ for $\langle X, X\rangle_t$

by computing a "bias corrected" estimator:

$$[X,X]_t^{BCR} = [X,X]_t - \frac{2}{3\sqrt{N}} \left(\widehat{\int_0^t v_s \sigma_s dX_s} \right)_n$$

where $\left(\widehat{\int_0^t v_s \sigma_s dX_s} \right)_n$ would be a consistent estimator (when $n \to \infty$) of $\int_0^t v_s \sigma_s dX_s$. An example of such a consistent estimator will be given in Section 5.1. Then, the new normalized estimation error $N^{1/2} \left([X,X]_t^{BCR} - \langle X,X \rangle_t \right)$ is asymptotically a mixture of normals whose mixture component is the variance coefficient equal to $\frac{2}{3} \int_0^t \left(u_s - \frac{2}{3} v_s^2 \right) \sigma_s^4 ds$. To understand that this variance coefficient is an efficiency bound, one may refer to the known result that an efficient Generalized Method of Moment estimator (GMM) reaches the semiparametric efficiency bound. More precisely, let us consider the simple case of estimation of the variance σ^2 of some random variable Z from an i.i.d. sample $Z_1, ..., Z_n$ when we have the extra information that the expectation $E(Z) = 0$. In this case, we know that a semi-parametrically efficient estimator of σ^2 is given by an efficient GMM estimator associated to the two moment conditions $\{E(Z) = 0, E(Z^2 - \sigma^2) = 0\}$ that is (see e.g. Back and Brown (1993)):

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

where $\hat{b}$ is the sample counterpart of the population regression coefficient:

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

In other words, we improve the naive estimator of $\text{Var}(Z) = E(Z^2)$ by computing a sample mean of the residual of the regression of Z^2 on Z. By doing so, we optimally minimize the variance among all possible unbiased estimators of σ^2 provided by sample counterparts of expectations $E\left(Z^2 - \beta Z\right) = \mathrm{Var}(Z)$ for any real number β. The resulting asymptotic variance of the normalized estimation error $n^{1/2}\left(\hat{\sigma}^2 - \sigma^2\right)$ is:

$$\mathrm{Var}(Z^2) - b^2\mathrm{Var}(Z) = (\kappa - 1 - \xi^2)[\mathrm{Var}(Z)]^2,$$

where $\kappa = E(Z^4)/[\mathrm{Var}(Z)]^2$ and $\xi = E(Z^3)/[\mathrm{Var}(Z)]^{3/2}$ respectively stand for the kurtosis and the skewness coefficient of the variable Z.

This strategy, well-known in Monte-Carlo estimation and dubbed the "control variable principle", is using efficiently the extra-information that $E(Z) = 0$. Of course, this principle is relevant only when the regression coefficient is not zero, that is, in the above example, when the variable Z features a non-zero skewness.

It is then clear that the estimator improvement allowed by Theorem 1 corresponds exactly to the same principle. The efficient asymptotic variance is defined from the integral between 0 and 1 of the function $\frac{2}{3}\left(u_s - \frac{2}{3}v_s^2\right)\sigma_s^4$ that is the local analog of $(\kappa - 1 - \xi^2)[\mathrm{Var}(Z)]^2$ (recall that in the normal case $u_s = \kappa = 3$ and $v_s = \xi = 0$). The deep reason for this analogy is the possibility, by following Mykland and Zhang (2006), to define in continuous time a regression of the naive normalized estimation error $n^{1/2}\left([X,X]_t - \langle X,X\rangle_t\right)$ on the process X_t, that is the continuous time analog of the regression above of Z^2 on Z. For doing so, we first note that by the Girsanov's theorem, we can make a measure change such that the process X_t is a local martingale under the new measure, dubbed equivalent martingale measure. This measure change, allowing to represent X_t as a process with zero-drift, is the continuous-time analog of the information that $E(Z) = 0$. The nice thing with continuous time is that we now have this information for free. Moreover, following Proposition 1 (p. 1408) of Mykland and Zhang (2009), we know that stable convergence in law is not impacted by this change of measure.

Then, under the equivalent martingale measure, the normalized estimation error is asymptotically equivalent to the local martingale

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

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

$$M_t = n^{1/2}\left(2\sum_{t_{n,i}\le t}\int_{t_{n,i}}^{t_{n,i+1}}\left(X_s - X_{t_{n,i}}\right)dX_s + 2\int_{t^*}^{t}(X_s - X_{t^*})dX_s\right).$$

The continuous time analog of the regression of the estimation error on the path of X_t amounts to characterize a stochastic process g_t solution of:

$$P\lim_{n\to\infty}\left\langle M_{.} - \int_0^{.} g_s dX_s, X\right\rangle_t = 0.$$

The solution of this equation is actually characterized by:

$$\int_0^t g_s\sigma_s^2\, ds = P\lim_{n\to\infty}\langle M, X\rangle_t = \frac{2}{3}\int_0^t \widetilde{v}_s\sigma_s^3\, ds,$$

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

$$g_s = \frac{2}{3}\widetilde{v}_s\sigma_s.$$

and the so-called bias term in Theorem 1 which should be (after rescaling by $1/\sqrt{n}$) subtracted from integrated variance to get an improved estimator is nothing but the regression:

$$\int_0^t g_s dX_s = \frac{2}{3}\int_0^t \widetilde{v}_s \sigma_s dX_s.$$

Hence, it is fair to say that our improved estimation strategy is the continuous time analog of the control variables principle. This principle 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.

We provide in the next section a list of possible models of random times, showing that some of them feature the kind of endogeneity that provides a non-zero tricity ($v_t \neq 0$) 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 $u_t = 3H_t'$. Equidistant sampling is a 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 accommodates some endogeneity of time, 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 (2007), 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 (2011) 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 (2007) 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 such that $t_{n,N} \leq 1 < t_{n,N+1}$.

In other words,

$$X_{t_{n,i+1}} - X_{t_{n,i}} = n^{-1/2} Z_{i+1} \text{ for } t_{n,i+1} < 1, \tag{13}$$

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 $N/n \xrightarrow{p} 1/(ab)$, and so the conditions of Theorem 1 are satisfied, with $v_t \equiv E(Z^3)/(ab)^{3/2} = (a-b)/\sqrt{ab}$ and $u_t \equiv E(Z^4)/(ab)^2 = (a^2 - ab + b^2)/ab$. We note that v_t is nonzero except when $a = b$. Moreover, it can be shown that the quadratic variation of time $H_t = t(a^2 + 3ab + b^2)/(3ab)$, hence it is never the case that $u_t = 3H'_t$. □

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 (13) 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 (2012) generalizes the hitting time technique of Example 4 above to construct autoregressive conditional duration models. It rests upon a dynamic version of Abbring (2012)'s mixed hitting time model. The observation times are defined recursively as:

$$t_{i+1} = \inf\{t > t_i : |Z_t - Z_{t_i}| > \varphi_{t_i} M_i\} \tag{14}$$

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 that are independent of Z. By contrast, the positive variable φ_{t_i} is $\mathcal{F}_{t_i}$-measurable and captures observed heterogeneity in the thresholds and associated hitting times. Given both $\mathcal{F}_{t_i}$ 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 $\mathcal{F}_{t_i}$ and M_i) denoted by $\rho_{t_i}(M_i)$. It can then be shown that conditionally on $\mathcal{F}_{t_i}$ and M_i, the log-price change $\Delta X_{t_{i+1}}$ has the same distribution as the following random variable

$$\begin{aligned}&[\mu_{t_i}(M_i) - \rho_{t_i}(M_i)\sigma_{t_i}(M_i)\mu_Z]\Delta t_{i+1}\\ &+ \rho_{t_i}(M_i)\sigma_{t_i}(M_i)\Delta Z_{t_{i+1}} + \sqrt{1-\rho_{t_i}(M_i)^2}\sigma_{\Delta t_{i+1}}(M_i)\sqrt{\Delta t_{i+1}}\zeta, \quad \text{where} \quad \zeta_i \sim N(0,1).\end{aligned}$$

In the simplest case when $\varphi_{t_i} \equiv 1$, $M_i \equiv M_n$, $\mu_{t_i}(M_i) \equiv \mu_X$, $\sigma_{t_i}(M_i) \equiv \sigma$, $\rho_{t_i}(M_i) \equiv \rho$ and $\mu_X = \rho\sigma\mu_Z$, $\Delta X_{t_{i+1}}$ are i.i.d. random variables which have the same distribution as the following random variable

$$\widetilde{X}_n := \rho\sigma Z_T + \sqrt{1-\rho^2}\sigma\sqrt{T}\zeta, \quad \text{where} \quad \zeta \sim N(0,1),$$

where $T = \inf\{t > 0 : |Z_t| > M_n\}$. It is easy to see that

$$E(\widetilde{X}_n^3) = \sigma^3\left(\rho^3 E(Z_T^3) + 3\rho(1-\rho^2)E(Z_T T)\right).$$

When $\mu_Z \neq 0$, one can show that $E(Z_T^3) \neq 0$ and $E(Z_T T) \neq 0$, hence as long as ρ is not the root of a cubic polynomial, the third moment $E(\widetilde{X}_n^3) \neq 0$. However, in order to obtain high frequency observations one needs to let $M_n \to 0$, say, for example, $M_n = M/\sqrt{n}$, then one can show that $\sqrt{n}^3 E(\widetilde{X}_n^3) \to 0$ and hence a nonzero v_t does not show up. [6] In the general case when φ_{t_i}, M_i etc. are nonconstant, a similar but more complicated computation still applies and v_t still vanishes. If however one chooses to use an asymmetric threshold in (15) like in Example 4 above, say, for example,

$$t_{i+1} = \inf\{t > t_i : Z_t - Z_{t_i} > \varphi_{t_i} M_i^1 \quad \text{or} \quad Z_t - Z_{t_i} < -\varphi_{t_i} M_i^2\}, \tag{15}$$

where for both $j = 1$ and 2, $\{M_i^j, i = 1, 2, ..\}$ is a sequence consisting of i.i.d. positive random variables, and the two sequences are independent and have different means, say with means $a/\sqrt{n}$ and $b/\sqrt{n}$ for some $a \neq b$ respectively, then a nonzero v_t will show up, just like in Example 4 above. □

Example 7. (Connection to uncertainty zones). Robert and Rosenbaum (2010, 2012) 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.

[6] We thank one of the referees for pointing this out.

In Robert and Rosenbaum (2012) there is no asymptotic bias (see their Lemma 4.14). This is not because of zero skewness, but rather, due to a cancelation of nonzero skewness terms, see Fukasawa and Rosenbaum (2012) for a detailed analysis. 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 (2010) 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

In this section we present a test for endogeneity of times, by establishing a CLT for the tricity $[X, X, X]_1$ under the null hypothesis that the observation times $t_{n,i}$ are independent of X_t.[7] We then apply the CLT to real data and compute the p-values. We shall see in Section 6 that when applied to the financial data that we consider, the test rejects the null hypothesis of non-endogeneity.

[7] Obviously, by a time change argument, the null hypothesis also covers times that are of the form $F(t_{n,i})$, where F is random, adapted, but independent of n. This is a form of endogeneity that does not lead to modification of the properties of the realized volatility.

4.1 CLT for tricity

THEOREM 2. *Assume the null hypothesis that the observation times $t_{n,i}$ are independent of X_t, that (1) holds, and that H_t admits a continuous derivative h_t. Suppose further that the following so-defined "tricity of time"*

$$Q_t := \lim_n N^2 \sum_{t_{n,i} \leq t} \Delta t_{n,i}^3, \quad t \in [0,1] \tag{16}$$

exists.[8] *Then, stably in law,*

$$N[X,X,X]_1 \to b + aZ, \tag{17}$$

where

$$b = 3\int_0^1 \sigma_t^2 h_t \; dX_t + \frac{3}{2}\int_0^1 h_t \; d\langle \sigma^2, X\rangle_t, \quad a^2 = 15\int_0^1 \sigma_t^6 \; dQ_t - 9\int_0^1 \sigma_t^6 h_t^2 \; dt,$$

and Z is a standard normal random variable that is independent of $\mathcal{F}_1$.

When the times are equidistant, the theorem is a special case of Theorem 2 in Kinnebrock and Podolskij (2008). See also the development in Example 3 of Mykland and Zhang (2009).

We next study how to estimate the parameters b and a.

For notational ease, we write $t_{n,i}$ as t_i in the rest of this section.

LEMMA 1. *Assume that $1/\min_i \Delta t_i = o_p(n^{1+\varepsilon})$ and $\max_i \Delta t_i = o_p(n^{-(1-\varepsilon)})$ for some $0 < \varepsilon < 1/2$. Then for any $1 > \beta \geq (1+\varepsilon) - 1/2$, if we let $M_n = [n^\beta]$, and define $\widehat{\sigma_t^2}$*

[8] Measurability of N is assured since the times are taken to be independent of the process X_t

as

$$\widehat{\sigma_t^2} = \frac{[X,X]_{t_{iM_n}} - [X,X]_{t_{(i-1)M_n}}}{t_{iM_n} - t_{(i-1)M_n}}, \qquad \text{if } t \in [t_{iM_n}, t_{(i+1)M_n}),\ i = 1,2,\ldots,[N/M_n], \quad (18)$$

and $\widehat{\sigma_t^2} = 0$ *for* $t \in [0, t_{M_n})$*. Then under the assumption that* σ_t *is continuous,* $\widehat{\sigma_t^2}$ *converges in probability in* $\mathbb{D}(0,1]$ *to* σ_t^2*.*

Observe that we intentionally shift the time when we define $\widehat{\sigma_t^2}$, so that it is adapted to the filtration $\mathcal{F}_t$.

Clearly if we define $\widehat{\sigma}_t := \sqrt{\widehat{\sigma_t^2}}$, then for any $\alpha > 0$, $\widehat{\sigma}_t^\alpha$ converges in probability in $\mathbb{D}(0,1]$ to σ_t^α. See Renò (2008), Kristensen (2010) for alternative estimators of spot volatility.

In the following we fix $M_n = [n^\beta]$ for some $1 > \beta \geq (1+\varepsilon)/(1-\varepsilon) - 1/2$.

We next estimate the derivatives of H_t and Q_t. For given n and $t \in (0,1)$, we set

$$\begin{cases} \widehat{H}_t = N \sum_{j \leq iM_n} \Delta t_j^2, \\ \widehat{Q}_t = N^2 \sum_{j \leq iM_n} \Delta t_j^3, \end{cases} \qquad \text{for } t = t_{iM_n} \text{ and piecewise linear in between the } t_{iM_n}.$$

The derivatives, $\widehat{h}_t$ and $\widehat{q}_t$, are then naturally defined as

$$\begin{cases} \widehat{h}_t = \dfrac{\hat{H}_{t_{iM_n}} - \hat{H}_{t_{(i-1)M_n}}}{t_{iM_n} - t_{(i-1)M_n}}, \\ \widehat{q}_t = \dfrac{\hat{Q}_{t_{iM_n}} - \hat{Q}_{t_{(i-1)M_n}}}{t_{iM_n} - t_{(i-1)M_n}}, \end{cases} \qquad \text{if } t \in [t_{iM_n}, t_{(i+1)M_n}),\ i = 1,2,\ldots, \quad (19)$$

and $\widehat{h}_t = \widehat{q}_t = 0$ for $t \in [0, t_{M_n})$. Again, here we shift the time to make them to be adapted to $\mathcal{F}_t$.

Proposition 1. *Under the assumptions of Theorem 2 and Lemma 1, the terms in the expressions of* b *and* a^2 *can be estimated as follows: let* $M_n = [n^\beta]$ *for some* $1 > \beta \geq (1+\varepsilon)/(1-\varepsilon) - 1/2$. *Then*

(i) the stochastic integral $(\widehat{\sigma}^2\widehat{h}) \cdot X$ *converges in probability to* $\int_0^1 \sigma_t^2 h_t \, dX_t$*;*

(ii) define

$$\widehat{\langle \sigma^2, X\rangle}_t = 2\sum_{j\leq i}(\widehat{\sigma^2_{t_{jM_n}}} - \widehat{\sigma^2_{t_{(j-1)M_n}}}) \cdot (X_{t_{jM_n}} - X_{t_{(j-1)M_n}}), \quad t \in [t_{iM_n}, t_{(i+1)M_n})),$$

and $\widehat{\langle \sigma^2, X\rangle}_t = 0$ *for* $t \in [0, t_{M_n})$. *Then* $\widehat{\langle \sigma^2, X\rangle}_t$ *converges in probability in* $\mathbb{D}[0,1]$ *to* $\langle \sigma^2, X\rangle_t$*;*

(iii) $\sum_i \widehat{h}_{t_{(i-1)M_n}} \cdot \left(\widehat{\langle \sigma^2, X\rangle}_{t_{iM_n}} - \widehat{\langle \sigma^2, X\rangle}_{t_{(i-1)M_n}}\right) \to \int_0^1 h_t \, d\langle \sigma^2, X\rangle_t$,
$\sum_i \widehat{\sigma}^6_{t_{(i-1)M_n}} \cdot \widehat{q}_{t_{(i-1)M_n}} \cdot (t_{iM_n} - t_{(i-1)M_n}) \to \int_0^1 \sigma_t^6 \, dQ_t$, *and*
$\sum_i \widehat{\sigma}^6_{t_{(i-1)M_n}} \cdot \widehat{h}^2_{t_{(i-1)M_n}} \cdot (t_{iM_n} - t_{(i-1)M_n}) \to \int_0^1 \sigma_t^6 h_t^2 \, dt$, *all in probability.*

We can then define estimators of b and a as follows

$$\begin{cases} \widehat{b} = 3(\widehat{\sigma}^2\widehat{h}) \cdot X + \dfrac{3}{2}\displaystyle\sum_i \widehat{h}_{t_{(i-1)M_n}} \cdot \left(\widehat{\langle \sigma^2, X\rangle}_{t_{iM_n}} - \widehat{\langle \sigma^2, X\rangle}_{t_{(i-1)M_n}}\right), \\ \widehat{a^2} = 15\displaystyle\sum_i \widehat{\sigma}^6_{t_{(i-1)M_n}} \cdot \widehat{q}_{t_{(i-1)M_n}} \cdot (t_{iM_n} - t_{(i-1)M_n}) - 9\sum_i \widehat{\sigma}^6_{t_{(i-1)M_n}} \cdot \widehat{h}^2_{t_{(i-1)M_n}} \cdot (t_{iM_n} - t_{(i-1)M_n}). \end{cases}$$

By the previous proposition, $\widehat{b}$ and $\widehat{a^2}$ converge in probability to b and a^2 respectively. Define $\widehat{a} = \sqrt{\widehat{a^2}}$. Combining this convergence with Theorem 2 we obtain

Corollary 1. *Under the assumptions of Proposition 1, for the afore-defined* $\widehat{b}$ *and* $\widehat{a}$ *we have that*

$$T := \frac{N[X, X, X]_1 - \widehat{b}}{\widehat{a}} \to Z,$$

where Z is a standard normal random variable that is independent of $\mathcal{F}_1$.

This result enables us to compute a p-value for the null hypotheses that the observation times are independent of the process X_t. More specifically, for each day, based on the observed price process X_{t_i} and the observation times t_i, we can compute the test statistic T. The asymptotic p-value is then given by $P(|Z| > |T|)$. Under the alternative (11) considered in this paper, $\sqrt{N}[X, X, X]_1$ has a non-zero limit, and thus we expect the statistic T to blow up asymptotically in this case. Hence the test should consistently detect the alternatives of non-zero limit of the tricity.

4.2 Combining Several Days

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 $(i = 1, \ldots, L)$ be the p-values from day 1 to day L, then under the null, asymptotically,

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

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

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

5 Simulation Study

5.1 Confidence Intervals in the Endogenous Case

We take the same setting as Example 4 in Section 3 with $\mu = 0$, $\sigma = 0.02$, $a = 0.04$, $b = 0.01$ and $n = 3600$. According to this stopping rule, a transaction happens each time when there is an increase of 0.0667% or a decrease of 0.0167% in the price.

We examine three confidence intervals based on the following three different methods.

- 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:

$$\sqrt{N}\left([X,X]_1 - \int_0^1 \sigma_t^2 dt\right) \to_{\mathcal{L}-Stably} \int_0^1 \sqrt{2\sigma_t^4 H'(t)}\, dB_t, \tag{20}$$

  where $B_t \perp\!\!\!\perp W_t$, and H_t is defined by (1).

- 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 quarticity which is equivalent to the above CLT if there were no endogeneity:

$$\sqrt{N}\left([X,X]_1 - \int_0^1 \sigma_t^2 dt\right) \to_{\mathcal{L}-Stably} \int_0^1 \sqrt{\frac{2}{3}u_s}\, \sigma_s^2\, dB_t, \tag{21}$$

  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 in Theorem 1, we use the blocking method as in Section 4.1. More specifically, we choose $\beta = 3/4$ and $M = [n^\beta]$. Moreover, for block number i covering the time period $(t_{(i-1)M}, t_{iM}]$ with length $\Delta\tau_i = t_{iM} - t_{(i-1)M}$, we estimate σ_t^2 by (18), estimate h_t and q_t by (19), and estimate v_t and u_t in a similar fashion. And so, for example, the asymptotic bias term $\frac{2}{3}\int_0^1 v_s\sigma_s \, dX_s$ when we build the CI_C based on Theorem 1 will be estimated by $\frac{2}{3}(\widehat{v}\cdot\widehat{\sigma})\cdot X$. Under the assumptions of Proposition 1, by Theorem VI.6.22 (c) of Jacod and Shiryaev (2003), this converges in probability to $\frac{2}{3}\int_0^1 v_s\sigma_s \, dX_s$.

Data of 252 days are simulated based on the parameters as listed above. Confidence intervals of the first 22 days are plotted in the upper panel of the Figure 1. Confidence intervals of 252 days are plotted in the bottom panel of Figure 1. The summary statistics comparing the performance of the confidence intervals based on the 252 days are listed in Table 1.

[Figure 1 about here]

[Table 1 about here]

From Figure 1 and Table 1 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 $\lim_n \frac{2}{3}N[X,X,X,X]_1$ may be substantially different from $\int_0^1 2\sigma_s^4 \, dH_s$, 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 meanwhile 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.

5.2 Confidence Intervals in the Non-endogenous Case

To further test the performance of the CI_C, we also simulated non-endogenous data. The X_t process is again taken to be σW_t for $\sigma = 0.02$ and W_t a Brownian motion. The observation times $t_{n,i}$'s are taken to be equidistant, namely, $t_{n,i} = i/n$ for $i = 0, 1, \ldots, n$ and $n = 3000$. The comparisons of the three different confidence intervals are summarized in Figure 2 and Table 2.

[Figure 2 about here]

[Table 2 about here]

We see from Figure 2 and Table 2 that for non-endogenous data, our CI_C performs just as well as the conventional confidence intervals.

6 Empirical Study

6.1 Data Description

We use trade data from the TAQ database. We consider several traded stocks at NYSE. Our analysis is based on subsampled log prices, with the sampling period K specified for different stocks and is reported below. The K is chosen to be reasonably large so that the microstructure noise would be negligible. The value n, which now characterizes the frequency of subsampled data, is estimated based on historical data and is reported below too. The block size M_n that is used in estimating $\widehat{\sigma_t^2}, \widehat{h}_t, \widehat{q}_t$ etc. is taken to be $[n^{3/4}]$.

We now conduct the test established in Section 4.1.

6.2 Test Results

We here study the behavior of our test statistic for two stocks: SKS and IBM (the test results for two other stocks (DDS and MAT) are similar and are put in a supplementary document which is available on our webpages). We show the distribution of daily p -values for SKS over one year or IMB over 3 months, along with a combined p-value (see Section 4.2 for the definition). It is clear from the results that the null hypothesis of non-endogeneity is rejected for these stocks 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.

6.2.1 SKS

In this part we apply the test to SKS 2005 one year data. The sampling frequency is $K = 8$, and n is taken to be 250.

Uncorrelated condition between the daily p-values is examined, and the ACF plot together with the histogram of the daily p-values are shown in Figure 3. The combined p-value is 0 to eight (numerically) significant digits.[9]

[Figure 3 about here]

6.2.2 IBM

Nextg we examine the IBM 2005 Jan-Mar three months' data. K and n are taken to be 25 and 250 respectively. The histogram of the daily p-values and the ACF plot are shown in Figure 4. The combined p-value is 0 to eight significant digits.

[Figure 4 about here]

7 Conclusion

We have established a central limit theorem for the realized volatility for general dependent times. We illustrate by simulation study how our theory can be used to obtain

[9] I.e., the p-value is smaller than 10^{-8}. The bound is due to the assessed numerical accuracy of our calculation.

correct interval estimates of the integrated volatility in the general endogenous time setting. We also show that the endogeneity can exist in financial data, using a test based on a central limit theorem for tricity.

Acknowledgements

We are very grateful to the editor and anonymous referees for their very valuable comments and suggestions.

REFERENCES

Abbring, J. H. (2012), "Mixed hitting-time models," *Econometrica*, 80, 783-819.

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.

Back, K., and Brown, D., (1993), "Implied probabilities in GMM estimators," *Econometrica*, 61, 971–975.

Barndorff-Nielsen, O.E., Graversen, S.E., J. Jacod, and Shephard, N. (2006), "Limit theorems for bipower variation in financial econometrics," *Econometric Theory*, 22, 677–719.

Barndorff-Nielsen, O.E., Graversen, S.E., J. Jacod, Podolskij, M. and Shephard, N. (2006), "A central limit theorem for realised power and bipower variations of continuous semimartingales," in *From Stochastic Calculus to Mathematical Finance, The Shiryaev Festschrift*, Yu. Kabanov and R. Liptser and J. Stoyanov, eds. (Springer-Verlag, Berlin), p. 33-69.

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.

— (2004), "Power and bipower variation with stochastic volatility and jumps (with discussion)," *Journal of Financial Econometrics*, 2, 1-48.

Christensen, K., Podolskij, M. and Vetter, M. (2011), "On covariation estimation for multivariate continuous Ito semimartingales with noise in non-synchronous observation schemes," Working paper.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap methods and their application*, vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge: Cambridge University Press, with 1 IBM-PC floppy disk (3.5 inch; HD).

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.

Fan, J. , Li, Y. and Yu, K. (2010), "Vast Volatility Matrix Estimation using High Frequency Data for Portfolio Selection", (to appear in *Journal of the American Statistical Association*).

Fukasawa, M. (2010a), "Central limit theorem for the realized volatility based on tick time sampling," *Finance and Stochastics*, 14 (2010), 209-233.

Fukasawa, M. (2010b), "Realized volatility with stochastic sampling," *Stochastic Processes and Their Applications*, 120, 829–552.

Fukasawa, M. and Rosenbaum, M.(2012), "Central Limit Theorems for Realized Volatility under Hitting Times of an Irregular Grid," *Stochastic Processes and Their Applications*, forthcoming.

Grammig, J. and Wellner, M. (2002), "Modeling the interdependence of volatility and inter-transaction duration processes," *Journal of Econometrics*, 106, 369–400.

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

Hayashi, T., Jacod, J., and Yoshida, N. (2011), "Irregular sampling and cenral limit theorems for power variations: the continuous case," *Ann. Inst. H. Poincare, Probab. Statis.,*, 47: 4, 1197-1218.

Hayashi, T. and Yoshida, N.(2005), "On Covariance Estimation of Non-synchronously Observed Diffusion Processes," *Bernoulli*, 11, 359-379.

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.

Jacod, J. and Shiryaev, A. (2003), *Limit theorems for stochastic processes*, Berlin, Springer-Verlag, 2nd ed.

Kinnebrock, S. and Podolskij, M. (2008), "A note on the central limit theorem for bipower variation of general functions," *Stochastic Processes and their Applications*, 118, 1056–1070.

Kristensen, D.(2010), "Nonparametric filtering of the realized spot volatility: a kernel-based approach," *Econometric Theory*, 26, 60–93.

Lee, S. and Mykland, P.A. (2008), "Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics," *Review of Financial Studies*, 21, 2535-2563.

Li, Y., and Zhang, Z., and Zheng, X.(2013), "Volatility Inference in the Presence of Both Endogenous Time and Microstructure Noise," to appear in a special issue of *Stochastic Processes and their Applications* (Rainer Dahlhaus, Jean Jacod, Per Mykland, and Nakahiro Yoshida, eds).

Mancini, C. (2001), "Disentangling the Jumps of the Diffusion in a Geometric Jumping Brownian Motion," *Giornale dell'Istituto Italiano degli Attuari*, LXIV, 19-47.

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

Mykland, P. A. (1994), "Bartlett type identities for martingales," *Annals of Statistics*, 22, 21–38.

Mykland, P.A., Shephard, N., and Sheppard, K. (2012), "Efficient and feasible inference for the components of financial variation using blocked multipower variation." Working paper, University of Oxford.

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

Mykland, P. A. and Zhang, L. (2009), "Inference for continuous semimartingales observed at high frequency," *Econometrica*, 77, 1403–1455.

Mykland, P. A. and Zhang, L. (2012), "The Econometrics of High Frequency Data," in *Statistical Methods for Stochastic Differential Equations*, M. Kessler, A. Lindner, and M. Sørensen, eds. (Chapman and Hall/CRC Press), p. 109-190.

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. (2012), "The Dynamic Mixed Hitting-Time Model for Multiple Transaction Prices and Times," *working paper*.

Renault, E. and Werker, B. J. (2011), "Causality effects in return volatility measures with random times," *Journal of Econometrics*, 60:1, 272-279.

Renò, R.(2008), "Nonparametric estimation of the diffusion coefficient of stochastic volatility models," *Econometric Theory*, 24, 1174–1206.

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

Robert, C. Y. and Rosenbaum, M. (2010), "On the Microstructural Hedging Error," *SIAM Journal of Financial Mathematics*, 1, 427-453.

Robert, C. Y. and Rosenbaum, M (2012), "Volatility and Covariation Estimation when Microstructure Noise and Trading Times are Endogenous," *Mathematical Finance*, 22 (1), 133–164.

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

Ross, S. (1996), *Stochastic Processes*, New York: Wiley, 2nd ed.

Wang, D.C. and Mykland, P.A. (2011), "The Estimation of Leverage Effect with High Frequency Data," Working paper, University of Oxford.

Xiu, D. (2010), "Quasi-Maximum Likelihood Estimation of Volatility WIth High Frequency Data," *Journal of Econometrics,* 159, 235-250.

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

Zhang, L. (2006), "Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach," *Bernoulli* 12, 1019-1043.

Zhang, L. (2011), "Estimating Covariation: Epps Effect and Microstructure Noise," *Journal of Econometrics,* 160, 33-47.

Zhang, L., Mykland, P.A. and Mykland and Aït-Sahalia, Y. (2005), "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data," *Journal of the American Statistical Association,* 100, 1394-1411.

A Appendix

A.1 Proof of Theorem 1

Proof. By assumption (7), $P(\max_t |t_{n,i+1} - t_{n,i}| \geq n^{-(2/3+\varepsilon)}) \to 0$. We first argue that from this, without loss of generality, we can assume that

$$\max_i |t_{n,i+1} - t_{n,i}| \leq n^{-(2/3+\varepsilon)} \quad \text{almost surely.} \tag{A.1}$$

To see this, construct new observation times given by $\tilde{t}_{n,0} = 0$ and then recursively $\tilde{t}_{n,i+1} = \min(t_{n,i+1}, \tilde{t}_{n,i} + n^{-(2/3+\varepsilon)})$. We obtain that $P(\tilde{t}_{n,i} = t_{n,i} \text{ for all } i) \to 1$ as $n \to \infty$. Thus the conditions of Theorem 1 remain satisfied (with the same limiting quantities), while (A.1) is also satisfied.

Next, because we shall prove stable convergence, and because of the local boundedness and that $\inf_{t\in(0,1]} \sigma_t > 0$, we can without loss of generality assume that $|\mu_t|$ is bounded by a nonrandom constant, and $0 < c < \sigma_t \leq \sigma^+$ for all $t \in (0, 1]$ for some nonrandom constants c and σ^+(see Section 2.4.5 of Mykland and Zhang (2012)). One can further suppress μ as in Section 2.2 (p.1407-1409) of Mykland and Zhang (2009), and act as if X is a 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 $t_{n,i}$ smaller than or equal to t. From (10), it follows as in the proof of Proposition 2 (p. 1952) of Mykland and Zhang (2006) that $\langle M, M\rangle_t \xrightarrow{p} \frac{2}{3}\int_0^t \widetilde{u}_s \sigma_s^4\, ds$ 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 (7). More 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

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

We shall show that $n\int_0^{\cdot}(X_t - X_{t_*})^3\, dX_t \to 0$ in $\mathbb{D}[0, 1]$. Now 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 $n\int_0^{\cdot}(X_t - X_{t_*})^3\, dX_t$

satisfies

$$
\begin{aligned}
&E\left\langle \int_0^{\cdot} n(X_t - X_{t_*})^3 \, dX_t, \int_0^{\cdot} n(X_t - X_{t_*})^3 \, dX_t \right\rangle_1 \\
&= n^2 E \int_0^1 (X_t - X_{t_*})^6 \, d\,\langle X, X\rangle_t \\
&\le c n^2 \sigma_+^8 \cdot E \int_0^1 (t - t_*)^3 \, dt \\
&\le c n^2 \sigma_+^8 \cdot n^{-(2/3+\varepsilon)\cdot 3} \\
&= O(n^{-3\varepsilon}) \to 0 \ ,
\end{aligned}
$$

where c is a universal constant, and where the second-to-last transition is by (A.1). This term is hence negligible. By Assumption (10) and (A.2), $\langle M, M\rangle_t \xrightarrow{p} \frac{2}{3}\int_0^t \widetilde{u}_s \sigma_s^4 ds$ for all t.

Similarly, (11) yields that $\langle X, M\rangle_t \xrightarrow{p} \frac{2}{3}\int_0^t \widetilde{v}_s \sigma_s^3 \, ds$, again for all t. In fact, by Itô's formula and that $d\langle X, M\rangle_t = 2n^{1/2}(X_t - X_{t_*}) \, d\langle X, X\rangle_t$ we get an analogous equation to (A.2) for $\langle X, M\rangle_t$ as follows:

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

Using similar arguments as in the previous paragraph, one can show that the martingale term is negligible. In fact, using (A.1) and the Burkholder-Davis-Gundy inequality again,

$$
\begin{aligned}
&E\left\langle 2n^{1/2} \int_0^{\cdot} (X_t - X_{t_*})^2 \, dX_t, 2n^{1/2} \int_0^{\cdot} (X_t - X_{t_*})^2 \, dX_t \right\rangle_1 \\
=&4nE \int_0^1 (X_t - X_{t_*})^2 \, d\langle X, X\rangle_t \\
\le&4n\sigma_+^4 E \int_0^1 (t - t_*)^2 \, dt \\
=&O(n^{-1/3-2\varepsilon}) \to 0.
\end{aligned}
$$

The convergence of $\langle X, M\rangle_t$ then follows from Assumption (11) and (A.3).

The overall result now follows from the limit results in either Theorem B.4 (p. 65-67) of Zhang (2001), or Theorem 2.28 of Mykland and Zhang (2012). A similar but less complicated argument was used (in the non-edogenous times case) in Proposition 1 and Theorem 1 (p. 1940-41) of Mykland and Zhang (2006). □

A.2 Proof of Remark 3

Proof. In analogy with Chapter 2.4.5 of Mykland and Zhang (2012), we can without loss of generality assume that σ_t^2 and $\langle \sigma^2, \sigma^2\rangle_t'$ are bounded by a constant, say, c_1.

Moreover, as argued in the proof of Theorem 1 we can assume (A.1), i.e., $\max_i \Delta t_{n,i} \leq n^{-(\frac{2}{3}+\varepsilon)}$.

Also, it is enough to show the result under the equivalent martingale measure discussed in Section 2.2 of Mykland and Zhang (2009). We here also assume that σ_t^2 is a martingale, which can in most cases be assumed through an additional Girsanov change-of-measure. (If this is not available, a direct but tedious calculation provides the same result). By the third Bartlett identity for martingales (eq. (2.14), p. 23, of Mykland (1994)), we have that

$$E((\Delta X_{t_{n,i}})^3 \mid \mathcal{F}_{t_{n,i-1}}) = 3E(\Delta X_{t_{n,i}} \int_{t_{n,i-1}}^{t_{n,i}} \sigma_s^2 ds \mid \mathcal{F}_{t_{n,i-1}}). \tag{A.4}$$

By Itô's formula, we obtain $d((t-t_{n,i-1})(\sigma_t^2-\sigma_{t_{n,i-1}}^2)) = (\sigma_t^2-\sigma_{t_{n,i-1}}^2)dt+(t-t_{n,i-1})d\sigma_t^2$, whence, from (A.4),

$$E((\Delta X_{t_{n,i}})^3 \mid \mathcal{F}_{t_{n,i-1}}) = 3\sigma_{t_{n,i-1}}^2 E(\Delta X_{t_{n,i}} \Delta t_{n,i} \mid \mathcal{F}_{t_{n,i-1}}) + \text{ remainder}, \tag{A.5}$$

where

$$\begin{aligned}\text{remainder } &= 3E(\Delta X_{t_{n,i}} \Delta t_{n,i} \Delta \sigma_{t_{n,i}}^2 \mid \mathcal{F}_{t_{n,i-1}}) - 3E(\Delta X_{t_{n,i}} \int_{t_{n,i-1}}^{t_{n,i}} (t-t_{n,i-1}) d\sigma_s^2 \mid \mathcal{F}_{t_{n,i-1}}) \\ &= 3E(\Delta X_{t_{n,i}} \Delta t_{n,i} \Delta \sigma_{t_{n,i}}^2 \mid \mathcal{F}_{t_{n,i-1}}) - 3E\left(\int_{t_{n,i-1}}^{t_{n,i}} (t-t_{n,i-1}) d\langle X, \sigma^2\rangle_{t_{n,i}} \mid \mathcal{F}_{t_{n,i-1}}\right).\end{aligned} \tag{A.6}$$

If we can show that the remainder (A.6) is of order $o_p(n^{-1/2})$, the result in Remark 3 follows since the difference between the optional (observed) and predictable variation is a martingale of higher order (compare, for example, to the proof in Chapter 2.3.6 of Mykland and Zhang (2012)). To see the negligibility of the remainder, focus on the first (more complicated) term; the second term is dealt with similarly but more easily.

By Hölder's inequality, and then the Burkholder-Davis-Gundy (BDG) inequality (see Section 3 of Ch. VII of Dellacherie and Meyer (1982), or p.193 and 222 in Protter (2004)), we obtain (C_4 is a universal constant in the BDG inequality):

$$\begin{aligned}&E(\Delta X_{t_{n,i}} \Delta t_{n,i} \Delta \sigma_{t_{n,i}}^2 \mid \mathcal{F}_{t_{n,i-1}}) && \text{(A.7)} \\ \leq& E((\Delta X_{t_{n,i}})^4 \mid \mathcal{F}_{t_{n,i-1}})^{\frac{1}{4}} E((\Delta t_{n,i})^2 \mid \mathcal{F}_{t_{n,i-1}})^{\frac{1}{2}} E((\Delta \sigma_{t_{n,i}}^2)^4 \mid \mathcal{F}_{t_{n,i-1}})^{\frac{1}{4}} \\ \leq& C_4^2 E((\Delta\langle X, X\rangle_{t_{n,i}})^2 \mid \mathcal{F}_{t_{n,i-1}})^{\frac{1}{4}} E((\Delta t_{n,i})^2 \mid \mathcal{F}_{t_{n,i-1}})^{\frac{1}{2}} E((\Delta\langle \sigma^2, \sigma^2\rangle_{t_{n,i}})^2 \mid \mathcal{F}_{t_{n,i-1}})^{\frac{1}{4}} \\ \leq& C_4^2 c_1 E((\Delta t_{n,i})^2 \mid \mathcal{F}_{t_{n,i-1}}). && \text{(A.8)}\end{aligned}$$

Hence,

$$\begin{aligned}\sum_i E(\Delta X_{t_{n,i}} \Delta t_{n,i} \Delta \sigma^2_{t_{n,i}} \mid \mathcal{F}_{t_{n,i-1}}) &\leq C_4^2 c_1 \sum_i E((\Delta t_{n,i})^2 \mid \mathcal{F}_{t_{n,i-1}}) \\ &\leq C_4^2 c_1 n^{-\frac{2}{3}-\varepsilon} \sum_i E(\Delta t_{n,i} \mid \mathcal{F}_{t_{n,i-1}}) \\ &= O_p(n^{-\frac{2}{3}-\varepsilon}), \end{aligned} \tag{A.9}$$

whence this term is negligible at the $o_p(n^{-\frac{1}{2}})$ rate. The result in the remark follows. □

A.3 Proof of Theorem 2

Proof. This is by direct modification of the proof in Example 3 of Mykland and Zhang (2009). In fact, equation (28) therein still holds. By (1) and (16) and similar calculations as on page 1415 of Mykland and Zhang (2009) we then obtain that, using the notation therein, under P_n^*, $N[X,X,X]_1$ converges to $b' + aN(0,1)$ where $b' = 3\int_0^1 \sigma_t^3 h_t \, dW_t^*$ and a as given in the theorem.

Next we translate the convergence from measure P_n^* back to P. By the same arguments as on p.1416 of Mykland and Zhang (2009) we have that

$$A_{12} = \frac{1}{2}\int_0^1 k_t \sigma_t^3 \, dH_t = \frac{3}{2}\int_0^1 \langle \sigma^2, X\rangle_t' \, dH_t = \frac{3}{2}\int_0^1 h_t \, d\langle \sigma^2, X\rangle_t.$$

The rest then follows from the same arguments as in Mykland and Zhang (2009). □

A.4 Proof of Lemma 1

Proof. No matter whether the observation times are endogenous or not, the convergence rate for $[X,X]_t$ is always $\sqrt{n}$, hence for our chosen M_n, if we let $\delta_n = \min_i(t_{iM_n} - t_{(i-1)M_n})$, then $1/\delta_n = o_p(n^{(1+\varepsilon)}/n^{\beta}) = o_p(\sqrt{n})$, and

$$1/\delta_n \cdot ([X,X]_t - \langle X,X\rangle_t) \;\to\; 0 \in \mathbb{D}([0,1]).$$

It then follows easily that $\widehat{\sigma_t^2}$ converges uniformly to σ_t^2 on any compact interval inside $(0,1]$, see, e.g., the proof of Lemma 1 in Li, Zhang, and Zheng (2013). □

A.5 Proof of Proposition 1

Proof. The conclusion in (i) follows from Lemma 1 and Theorem VI.6.22 (c) of Jacod and Shiryaev (2003); (ii) follows as in Section 4.3 in Mykland and Zhang (2009) along with the development in Wang and Myland (2011); (iii) is straightforward. □

A.6 Figures and Tables

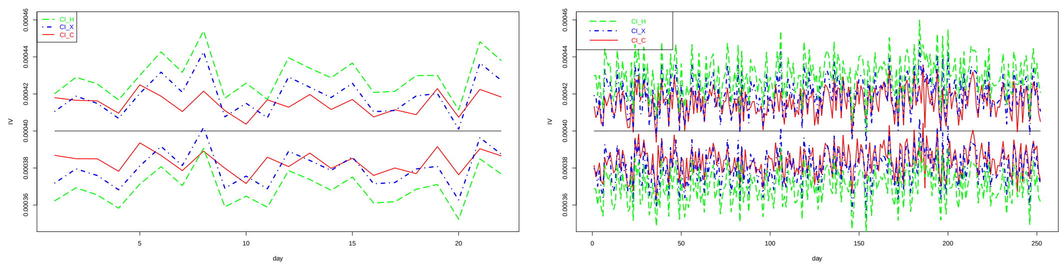

Figure 1. Confidence intervals computed based on the three methods described in Section 5.1 for endogenous data (Green dashed: CI_H; Blue dotted: CI_X; Red solid: CI_C). Left panel: for 22 days; right panel: for 252 days.

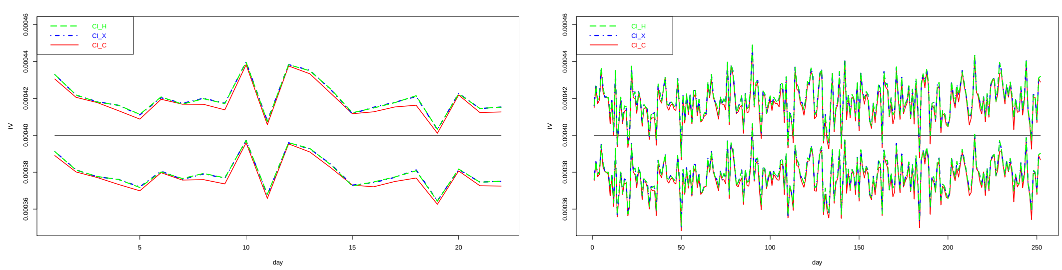

Figure 2. Confidence intervals computed based on the three methods described in Section 5.1 for non-endogenous data (Green dashed: CI_H; Blue dotted: CI_X; Red solid: CI_C). Left panel: for 22 days; right panel: for 252 days. Three different CIs roughly overlap each other.

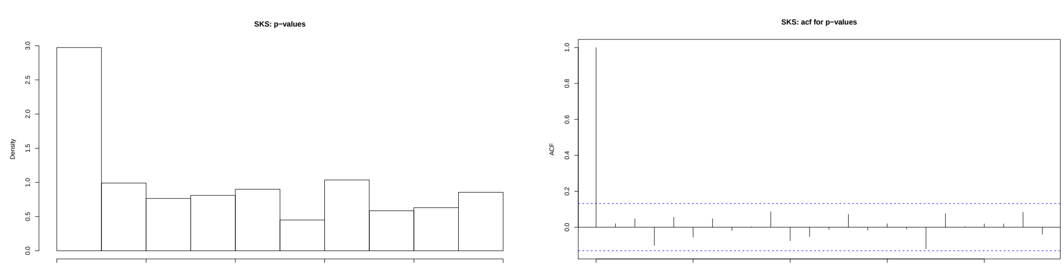

Figure 3. Histogram and ACF plot of the daily p-values for SKS 2005 one year data.

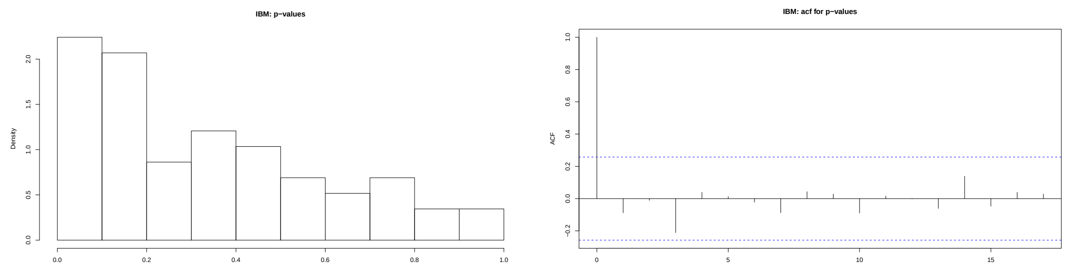

Figure 4. Histogram and ACF plot of the daily p-values for IBM 2005 Jan-Mar three months data.

	Average width	RMSE	Coverage Frequency	% Reduced width compared with CI_H	% Reduced RMSE compared with CI_H
CI_H	6.034e-05	1.001e-05	99.6%	–	–
CI_X	3.937e-05	1.001e-05	96.0%	34.7%	0
CI_C	3.165e-05	7.702e-06	95.5%	47.5%	23.1%

Table 1. Summary statistics based on simulated endogenous data of 252 days. 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. CI_H and CI_X are based on existing CLTs when endogeneity is ignored, see (20) and (21); CI_C is based on Theorem 1, which is the correct confidence interval to use for endogenous data.

	Average width	RMSE	Coverage Frequency
CI_H	4.045e-05	1.003e-05	93.7%
CI_X	4.041e-05	1.003e-05	93.6%
CI_C	4.025e-05	1.061e-05	93.4%

Table 2. Summary statistics based on simulated non-endogenous data of 252 days. The confidence intervals built based on the three methods are roughly the same.