Communications for Statistical Applications and Methods
2016, Vol. 23, No. 5, 411–421

http://dx.doi.org/10.5351/CSAM.2016.23.5.411
Print ISSN 2287-7843 / Online ISSN 2383-4757

Asymptotics for realized covariance under market microstructure noise and sampling frequency determination

Dong Wan Shin[a], Eunju Hwang[1,b]

[a]Department of Statistics, Ewha Womans University, Korea;
[b]Department of Applied Statistics, Gachon University, Korea

Abstract

Large frequency limiting distributions of two errors in realized covariance are investigated under noisy and non-synchronous high frequency sampling situations. The first distribution characterizes increased variance of the realized covariance due to noise for large frequency and the second distribution characterizes decreased variance of the realized covariance due to discretization for large frequency. The distribution of the combined error enables us to determine the sampling frequency which depends on a nuisance parameter. A consistent estimator of the nuisance parameter is proposed.

Keywords: market microstructure noise, non-synchronous trading, realized covariance

1. Introduction

During the recent decades there have been active research on statistical inference for integrated volatilities of financial asset returns using high-frequency data sets. Ultra-high frequency samples are known to be subject to market microstructure noises due to irregular trading, discreteness of prices, bid/ask bounce. The presence of market microstructure noise complicates volatility estimation, which causes some statistically serious problems such as inefficiency, bias, and inconsistency. Various methods for integrated volatilities under high frequency sampling were developed by Aït-Sahalia *et al.* (2005), Zhang *et al.* (2005) and Bandi and Russell (2008), and others.

Recent literature indicates significant efforts to estimate the integrated covariance of multiple assets. Barndorff-Nielsen and Shephard (2004) developed a general asymptotic theory for realized covariations such as covariance, regression coefficient, and correlation coefficient under fixed sampling and no-noise. Under high frequency sampling (in addition to noise) there is another factor of nonsynchronous trading which makes the efficient estimation of integrated covariance difficult as pointed out indicated by Hayashi and Yoshida (2005). Various attempts have been made to overcome the difficulty from the two factors of noise and nonsynchronousity as well as construct consistent and efficient realized covariances by Voev and Lunde (2007) for a subsampling method; Aït-Sahalia *et al.* (2010) via quasi-maximum likelihood estimator; Griffin and Oomen (2011) for comparison of several methods; Bibinger (2011a, 2012) via a generalized multi-scale method; Barndorff-Nielsen *et al.* (2011) via a multivariate realized kernel; Dovonon *et al.* (2013) via the i.i.d. bootstrapping, and Hwang and Shin (2016) via the stationary bootstrapping.

[1] Corresponding author: Department of Applied Statistics, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam 13120, Korea. E-mail: ehwang@gachon.ac.kr

Published 30 September 2016 / journal homepage: http://csam.or.kr

In this paper, we develop asymptotic distributions for approximation errors in the realized covariance in the presence of noise and nonsynchronousity. First, we formulate two approximations for errors in the realized covariance as an estimator of the quadratic covariance: one due to the noise and the other due to the discretization error of the latent process. Second, we develop a normal approximation of the realized covariance to the target integrated covariance from the approximations of the two errors. The normal approximation enables us to investigate a trade-off between two errors to reduce the variance that causes the inefficiency problem. The sample frequency will be optimized by minimizing the variance of the total error which depends on a nuisance parameter. We provide a consistent estimator for the nuisance parameter.

The remainder of the paper is organized as follows. In Section 2 we describe the preliminary setup and assumptions. Asymptotic results and optimal sampling are presented in Section 3 and the estimation for the asymptotic variance which determines the optimal sampling is discussed in Section 4. Proofs are given in Section 5.

2. Preliminary setup

This section describes a realized covariance based on high frequency bivariate nonsynchronous noisy samples from a two-dimensional diffusion for a couple of asset prices $X(t) = \{(X_1(t), X_2(t))' : t \geq 0\}$. The latent log-price is assumed to follow a continuous-time diffusion model

$$dX(t) = \mu(t)dt + \sigma(t)dB(t), \tag{2.1}$$

where $\mu(t) = (\mu_1(t), \mu_2(t))'$ is the drift vector, $\sigma(t) = [\begin{smallmatrix}\sigma_{11}(t) & \sigma_{12}(t)\\ \sigma_{21}(t) & \sigma_{22}(t)\end{smallmatrix}]$ is 2×2 volatility matrix, and $B(t) = (B_1(t), B_2(t))'$ is the standard 2-dimensional Brownian motions. Let $Y(t) = \{(Y_1(t), Y_2(t))' : t \geq 0\}$ denote the observable log-price process.

In ultra high-frequency sampling, two log-prices are observed asynchronously with market microstructure noise. During time interval $[0, T]$, the i^{th} asset is observed at times $t_{i\ell}$, $\ell = 1, 2, \ldots, N_i$, where N_i is the sample size of the i^{th} asset data set, $i = 1, 2$. Due to the non-synchronicity, typically $\{t_{1\ell}\} \neq \{t_{2\ell}\}$. Usually the high-frequency data are contaminated with noise:

$$Y_i(t_{i\ell}) = X_i(t_{i\ell}) + \epsilon_i(t_{i\ell}), \quad \ell = 1, \ldots, N_i, \; i = 1, 2. \tag{2.2}$$

The noise processes are assumed to satisfy the following assumptions:

(A1) For $i = 1, 2$, the noises $\{\epsilon_i(t_{i\ell}), \ell = 1, \ldots, N_i\}$ are i.i.d. random variables with mean zero, finite variance $E\epsilon_i^2 < \infty$, finite fourth moment $E\epsilon_i^4 < \infty$. The noise processes $\{\epsilon_i(t)\}, i = 1, 2$ are independent of each other and independent of the processes $\{X_i(\cdot), i = 1, 2\}$.

The integrated covariance over a fixed time interval $[0, T]$ is defined by

$$\langle X_1, X_2 \rangle_T = \int_0^T \Sigma_{12}(t)dt, \quad \Sigma_{12}(t) = (\sigma(t)\sigma(t)')_{12}.$$

Note that $\Sigma_{12}(t) = \sigma_{11}(t)\sigma_{21}(t) + \sigma_{12}(t)\sigma_{22}(t) = \rho(t)\sigma_1(t)\sigma_2(t)$ where $\sigma_i^2(t) = \Sigma_{ii}(t) = (\sigma(t)\sigma(t)')_{ii} = \sigma_{i1}^2(t) + \sigma_{i2}^2(t)$, $i = 1, 2$ and $\rho(t) = \Sigma_{12}(t)/\sqrt{\Sigma_{11}\Sigma_{22}(t)}$.

As an estimator of the integrated covariance, a realized covariance is constructed from a nonsynchronous sample. To handle the asynchronous observations of the two assets, we consider the synchronizing way of Barndorff-Nielsen *et al.* (2011), called the *refresh time*. The refresh times equal to the *closest synchronous approximation* of Bibinger (2012, p.2418). See Figure 1 of Barndorff-Nielsen *et al.* (2011) for an illustration of refresh times.

Refresh time is described briefly. The first refresh time τ_1 is the first time t before which both asset prices are observed; for each $k = 1, 2, \ldots$, the $(k+1)^{th}$ refresh time τ_{k+1} is the first time after τ_k before which both asset prices are observed. More formal definition follows. For $i = 1, 2$, let $N_i(t)$ be the number of observations in the i^{th} asset made up to time $t \in [0, T]$. Now we define the refresh time as: the first refresh time τ_1 is defined as $\tau_1 = \max(t_{11}, t_{21})$, and then subsequent refresh times are defined as, for $k = 1, 2, \ldots$, $\tau_{k+1} = \max(t_{1,N_1(\tau_k)+1}, t_{2,N_2(\tau_k)+1})$. Denote the resulting refresh time sample size by N, and assume $N = O_p((N_1 + N_2)/2) \to \infty$ as $(N_1 + N_2)/2 \to \infty$.

We rewrite the refresh times with N as $\{\tau_{N,1}, \tau_{N,2}, \ldots, \tau_{N,N}\}$. For each j, we denote the observed log-price observation at refresh time $\tau_{N,j}$ by $Y(\tau_{N,j}) \equiv (Y_1(\tau_{N,j}), Y_2(\tau_{N,j}))'$ which consists of new price at the refresh time $\tau_{N,j}$, say, $Y_1(\tau_{N,j})$, and of other price, say, $Y_2(\tau_{N,j})$ defined by the last observation of $Y_2(\cdot)$ traded in time interval $(\tau_{N,j-1}, \tau_{N,j}]$. The refresh time is subject to "stale pricing errors" in that only $Y_1(\tau_{N,j})$, say, is observed and the other $Y_2(\tau_{N,j})$ is refreshed rather than observed: $Y_2(\tau_{N,j})$ is a stale price rather than an observed price. According to Barndorff-Nielsen *et al.* (2011), under their assumptions of (A2) below, these stale pricing errors have no impact on the asymptotic distribution of estimators. We adopt the conditions of Barndorff-Nielsen *et al.* (2011).

Note that N is random. Let $\Delta_{N,j} := \tau_{N,j} - \tau_{N,j-1}$, $D_{N,j} := N\Delta_{N,j}$, for $j = 1, \ldots, N$, and let $\mathcal{F}_t$ be some filtration so that $X = (X_1, X_2)'$ is defined on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$. Denote $\lfloor x \rfloor$ by integer part of x. We assume

(A2) (i) $E(D^r_{N,\lfloor tN \rfloor} | \mathcal{F}_{\tau_{N,\lfloor tN \rfloor - 1}}) \xrightarrow{\text{p}} \kappa_r(t)$, $0 < r \leq 2$ as $N \to \infty$ where $\kappa_r(t)$ are strictly positive cádlág processes adopted $\{\mathcal{F}_t\}$; (ii) $\max_{j \in \{i+1, \ldots, i+R\}} D_{N,j} = o_p(R^{1/2})$ for any i; (iii) $\tau_{N,0} = 0$ and $\tau_{N,N+1} \geq T$.

Let $\mathcal{G} := \{\tau_{N,0}, \tau_{N,1}, \tau_{N,2}, \ldots, \tau_{N,N}\}$, with $\tau_{N,0} = 0$, be the full grid of time points that have been synchronized by the refresh time and $\mathcal{H} \subseteq \mathcal{G}$ be a subgrid denoted by $\mathcal{H} = \{t_0, t_1, \ldots, t_n\}$ where n is the number of time increments $(t_j, t_{j+1}]$ with $t_j, t_{j+1} \in \mathcal{H}$ and $n < N$, satisfying

$$\max_{t_j, t_{j+1} \in \mathcal{H}} \left(t_{j+1} - t_j\right) = O\left(\frac{1}{n}\right). \tag{2.3}$$

We investigate the limiting distribution of the errors in realized covariance based on samples on $\mathcal{H}$ as an estimator of the integrated covariance for large n in Section 3 below. The asymptotic result is useful to determine n, as shown in Section 3 and Section 4. Once n is determined, then the set of subsample time points $\mathcal{H} = \{t_0, t_1, \ldots, t_n\}$ is constructed as follows. Let $q = \lfloor N/n \rfloor$, the integer part of N/n. Each element t_j in $\mathcal{H}$ is chosen as $t_j = \tau_{N,qj}$ for $j = 1, 2, \ldots, n$, with $t_0 = 0$.

Realized covariance based on the subgrid $\mathcal{H}$ is defined as the quadratic covariation:

$$[Y_1, Y_2]_T \equiv [Y_1, Y_2]_T^{(\mathcal{H})} := \sum_{j=0}^{n-1} \left(Y_1(t_{j+1}) - Y_1(t_j)\right)\left(Y_2(t_{j+1}) - Y_2(t_j)\right).$$

The quadratic covariations like $[X_1, X_2]_T$, $[\epsilon_1, \epsilon_2]_T$, $[X_i, \epsilon_j]_T$, $i, j = 1, 2$, are defined in the same way.

Now for the discretization of X-process, we describe the quadratic variation of time. Let $H(t)$ be the asymptotic quadratic variation of time, as discussed by Mykland and Zhang (2002),

$$H(t) = \lim_{n \to \infty} \frac{n}{T} \sum_{t_j, t_{j+1} \in \mathcal{H}, t_{j+1} \leq t} \left(t_{j+1} - t_j\right)^2.$$

$H(\cdot)$ is well-defined under conditions of (2.3) above and of (A3) below, according to Proposition 1 on p. 1399, of Zhang *et al.* (2005, Section 2.3), (A3) below states a technical condition on the filtration $(\mathcal{F}_t)_{0\le t\le T}$ to which $X(t)$ and $\mu(t)$ are assumed to be adapted. We assume

(A3) (Description of the filtration) There is a continuous multidimensional P-local martingale $\mathcal{X} = (\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(p)})$ for any p, so that $\mathcal{F}_t$ is the smallest σ-field containing $\sigma\{\mathcal{X}_s, s \le t\}$ and $\mathcal{N}$, where $\mathcal{N}$ contains all of the null sets in $\sigma\{\mathcal{X}_s, s \le T\}$. For example, $\mathcal{X}$ can be a collection of Brownian motion.

3. Asymptotic results

We investigate the distributions of two errors in the realized covariance. The first error is associated with market microstructure noise and the second error is related with discretization. The error distribution in Lemma 1 below shows the increased variance of the realized covariance due to the noise for large n. The error distribution in Lemma 2 below reveals decreased error variance due to discretazation for large n. From the two lemmas, we develop a normal approximation of the realized covariance to the target integrated covariance in Theorem 1 below. The theorem shows that a trade-off between two errors is needed to reduce variance of the realized covariance.

Lemma 1. *(Noise error) We consider model (2.1) with noise (2.2) and assume (A1)–(A3) above. Then as $n \to \infty$ we have*

$$[Y_1, Y_2]_T = [X_1, X_2]_T + \varpi_n Z_{noise} + e_n + O_p\left(\frac{1}{\sqrt{n}}\right),$$

where

$$e_n := \epsilon_1(t_0)\epsilon_2(t_0) + \epsilon_1(t_n)\epsilon_2(t_n),$$
$$\varpi_n := \left(6nE\epsilon_1^2 E\epsilon_2^2 + 2E\epsilon_1^2[X_2, X_2]_T + 2E\epsilon_2^2[X_1, X_1]_T\right)^{\frac{1}{2}}$$

and Z_{noise} is a standard normal random variable.

According to Lemma 1, we see that the error due to the noise, $[Y_1, Y_2]_T - [X_1, X_2]_T$, is composed of two terms $\varpi_n Z_{noise}$ and e_n. The first term corresponds to inflated variance due to noise. The second term is the bias due to noises for the first and the last observations for the subgrid $\mathcal{H}$. Note that $e_n = O_p(\sqrt{E\epsilon_1^2 E\epsilon_2^2})$ and is not related with sampling schemes. It is interesting to compare this order-1 bias e_n for realized covariance with the order n bias $nE\epsilon_1^2$ of Zhang *et al.* (2005, Lemma 1) for the error $[Y_1, Y_1]_T - [X_1, X_1]_T$ of realized variance. The bias e_n is usually negligible because $E\epsilon_i^2$ are small in practice.

Therefore, $[Y_1, Y_2]_T$ almost unbiasedly estimates $[X_1, X_2]_T$. The variance ϖ_n increases at order of n as n increases. Lemma 1 implies that the smaller sample size makes the smaller error of the approximation of the unbiased estimator $[Y_1, Y_2]_T$ to the quadratic co-variation $[X_1, X_2]_T$, and thus the optimal choice of n is to make it as small as possible to reduce the error if we consider only inflated variance due to noise. However, we need to consider the inflated variance from the discretization error as well which increases as n decreases as shown in Lemma 2.

Lemma 2. *(Discretization error) We consider the model (2.1) with the noise (2.2) and assume (A1)–(A3) above. Then as $n \to \infty$ we have*

$$\left(\frac{n}{T}\right)^{\frac{1}{2}} ([X_1, X_2]_T - \langle X_1, X_2 \rangle_T) \xrightarrow{\text{d}} \vartheta Z_{discrete},$$

where

$$\vartheta := \left(\int_0^T \left[1 + \rho^2(t)\right] \sigma_1^2(t)\sigma_2^2(t)H'(t)\, dt \right)^{\frac{1}{2}}$$

and $Z_{discrete}$ is a standard normal random variable.

Now we combine Lemma 1 with Lemma 2 to establish the approximation of the realized covariance to the target integrated covariance. We see in Theorem 1 below that the total error $[Y_1, Y_2]_T - \langle X_1, X_2 \rangle_T$ of the realized covariance is given by the sum of two errors in Lemmas 1 and 2, of which one is increasing and the other is decreasing as $n \to \infty$.

Theorem 1. *We consider the model (2.1) with the noise (2.2) and assume (A1)–(A3) above. Then as $n \to \infty$ we have*

$$[Y_1, Y_2]_T = \langle X_1, X_2 \rangle_T + \upsilon_n Z_{total} + e_n + O_p\left(\frac{1}{\sqrt{n}}\right),$$

where Z_{total} is a standard normal random variable and $\upsilon_n := (\varpi_n^2 + (T/n)\vartheta^2)^{1/2}$, that is,

$$\upsilon_n^2 = 6nE\epsilon_1^2 E\epsilon_2^2 + 2E\epsilon_1^2[X_2, X_2]_T + 2E\epsilon_2^2[X_1, X_1]_T + \frac{T}{n}\int_0^T \left[1 + \rho^2(t)\right] \sigma_1^2(t)\sigma_2^2(t)H'(t)dt.$$

Remark 1. **(Optimal sampling frequency)** An optimal sampling frequency can be obtained by minimizing the total variance $v_n^2 = (\varpi_n^2 + (T/n)\vartheta^2)$. The optimal trade-off can be obtained from $\partial\upsilon_n^2/\partial n = 0$ as given by

$$n^* = \left(\frac{T\vartheta^2}{6E\epsilon_1^2 E\epsilon_2^2}\right)^{\frac{1}{2}}. \tag{3.1}$$

For this optimal n^*, the minimum variance is

$$\upsilon_{\min}^2 = 2\left(6TE\epsilon_1^2 E\epsilon_2^2 \vartheta^2\right)^{\frac{1}{2}} + 2E\epsilon_1^2[X_2, X_2]_T + 2E\epsilon_2^2[X_1, X_1]_T.$$

Remark 2. **(A feasible optimal sampling frequency)** The optimal n^* in (3.1) is not feasible because of nuisance parameters. For the nuisance parameters $E\epsilon_i^2$, $i = 1, 2$, we use

$$\widehat{E\epsilon_i^2} = \frac{1}{2N}[Y_i, Y_i]^{(all)}, \quad [Y_i, Y_i]^{(all)} = \sum_{\tau_j \in \mathcal{G}} \left(Y_i(\tau_j) - Y_i(\tau_{j-1})\right)^2, \ i = 1, 2,$$

which are consistent according to Zhang *et al.* (2005, Section 2.2). A consistent estimator $\widehat{\vartheta^2}$ of the remaining nuisance parameter ϑ^2 is constructed in Section 4 below from which we construct a feasible optimal sample size $\hat{n}^*$ given by

$$\hat{n}^* = \left(\frac{T\widehat{\vartheta^2}}{6\widehat{E\epsilon_1^2}\widehat{E\epsilon_2^2}}\right)^{\frac{1}{2}}.$$

4. A consistent estimator of ϑ^2

In order to construct a consistent estimator of ϑ^2, we follow a sub-grid method similar to Zhang *et al.* (2005, Sections 5 and 6) and apply an asymptotic result of Hwang and Shin (2016). Let an integer K depending on N be given. For $k = 1, \ldots, K$, let $\mathcal{G}^{(k)}$ be nonoverlapping subgrids of the full grid $\mathcal{G}$ with $\mathcal{G} = \bigcup_{k=1}^{K} \mathcal{G}^{(k)}$. A natural way to select $\mathcal{G}^{(k)}$ can be seen in Section 3.2 of Zhang *et al.* (2005) or in Section 2 of Hwang and Shin (2016), which is given by $\mathcal{G}^{(k)} = \{\tau_{N,(j-1)K+(k-1)} : j = 1, \ldots, m\}$ for $k = 1, \ldots, K$ where $m = [N/K]$. For example, if $N = 100$, $K = 20$, then we have $m = 5$, $\mathcal{G}^{(1)} = \{\tau_0, \tau_{0.2N}, \tau_{0.4N}, \tau_{0.6N}, \tau_{0.8N}, \tau_N\}, \mathcal{G}^{(2)} = \{\tau_1, \tau_{0.2N+1}, \tau_{0.4N+1}, \tau_{0.6N+1}, \tau_{0.8N+1}\}$, $\mathcal{G}^{(3)} = \{\tau_2, \tau_{0.2N+2}, \tau_{0.4N+2}, \tau_{0.6N+2}, \tau_{0.8N+2}\}, \ldots, \mathcal{G}^{(20)} = \{\tau_{19}, \tau_{0.2N+19}, \tau_{0.4N+19}, \tau_{0.6N+19}, \tau_{0.8N+19}\}$, where $\tau_{N,j}$ is denoted as τ_j. Note that $\mathcal{G}^{(k)}$ consists of every K^{th} point of $\mathcal{G}$ starting with the k^{th} point.

Let

$$[Y_1, Y_2]_t^{(k)} = \sum_{t_j, t_{j,+} \in \mathcal{G}^{(k)}, t_{j,+} \le t} \Big(Y_1(t_{j,+}) - Y_1(t_j)\Big)\Big(Y_2(t_{j,+}) - Y_2(t_j)\Big),$$

where $t_{j,+}$ is the following element of t_j in $\mathcal{G}^{(k)}$. Then $[Y_1, Y_2]_t^{(k)}$ is the realized covariance on the grid $\mathcal{G}^{(k)}$ up to time t. Let $[Y_1, Y_2]_T^{(avg)} = (1/K) \sum_{k=1}^{K} [Y_1, Y_2]_T^{(k)}$. By Theorem 3.1 of Hwang and Shin (2016), if $K = cN^{2/3}$ then we have

$$N^{\frac{1}{6}} \left([Y_1, Y_2]_T^{(avg)} - \langle X_1, X_2 \rangle_T\right) \xrightarrow{\text{d}} N\left(0, \varsigma^2\right), \tag{4.1}$$

where

$$\varsigma^2 = \frac{6}{c^2} E\epsilon_1^2 E\epsilon_2^2 + cT\vartheta^2. \tag{4.2}$$

The second term ϑ^2 of the asymptotic variance has been explicitly computed in Proposition A.1 of Bibinger (2011b).

As in Section 6 of Zhang *et al.* (2005), we consider a partition $[0, T_1], (T_1, T_2], \ldots, (T_{M-1}, T_M]$ of $[0, T]$ for some M. Note that $\langle X_1, X_2 \rangle_{T_m} - \langle X_1, X_2 \rangle_{T_{m-1}} = \int_{T_{m-1}}^{T_m} \Sigma_{12}(t)dt$ for $m = 1, 2, \ldots, M$, and its estimator is $[Y_1, Y_2]_{T_m}^{(avg)} - [Y_1, Y_2]_{T_{m-1}}^{(avg)}$.

For $m = 1, 2, \ldots, M$, let $T_m = (m/M)T$. First we focus on the m^{th} time period $[T_{m-1}, T_m]$ and apply the normality of (4.1) to this time interval. Let N_m be the number of points in the m^{th} time interval; $N = \sum_{m=1}^{M} N_m$. We may assume $N_m \to \infty$ as $N \to \infty$ for each m. Then, if $K_m = c_m N_m^{2/3}$, we have

$$N_m^{\frac{1}{6}} \left([Y_1, Y_2]_{T_m}^{(avg)} - [Y_1, Y_2]_{T_{m-1}}^{(avg)} - \int_{T_{m-1}}^{T_m} \Sigma_{12}(t)dt\right) \xrightarrow{\text{d}} N\left(0, \varsigma_m^2\right), \tag{4.3}$$

where

$$\varsigma_m^2 = \frac{6}{c_m^2} E\epsilon_1^2 E\epsilon_2^2 + c_m \frac{T}{M} \int_{T_{m-1}}^{T_m} \left[1 + \rho^2(t)\right] \sigma_1^2(t)\sigma_2^2(t)H'(t)dt. \tag{4.4}$$

The asymptotic results (4.3)–(4.4) enable us to construct a consistent estimator of ϑ^2 as follows. Similarly to $\mathcal{G} = \bigcup_{k=1}^{K} \mathcal{G}^{(k)}$, for each $k = 1, \ldots, K$, nonoverlapping subgrids of $\mathcal{G}^{(k)}$ are constructed by $\mathcal{G}^{(k)} = \bigcup_{i=1}^{I} \mathcal{G}^{(k,i)}$ for some number I, (which is less than the number of elements in $\mathcal{G}^{(k)}$), where $\mathcal{G}^{(k,i)}$ contains every I^{th} point of $\mathcal{G}^{(k)}$, starting with the i^{th} point. Then

$$\mathcal{G} = \bigcup_{k=1}^{K} \mathcal{G}^{(k)} = \bigcup_{k=1}^{K} \bigcup_{i=1}^{I} \mathcal{G}^{(k,i)}. \tag{4.5}$$

In order to define the estimator, we need a sequence of positive numbers $\{b_j : j = 1, 2, \dots\}$ such that $\lim_{N\to\infty} (1/N) \sum_{j=1}^{N} b_j = 1$ for which we choose $b_j = (j^2 - 1)/j^2$. Let $K^{(1)}_{m,j} = b_j M^{1/2} N_m^{2/3}$, $K^{(2)}_{m,j} = M^{1/2} K^{(1)}_{m,j} = b_j M N_m^{2/3}$, and $I_m = M^{1/2}$. Let

$$[Y_1, Y_2]^{(k)}_{(m)} = \sum_{t_j, t_{j,+} \in \mathcal{G}^{(k)} \cap [T_{m-1}, T_m]} \Big(Y_1(t_{j,+}) - Y_1(t_j)\Big)\Big(Y_2(t_{j,+}) - Y_2(t_j)\Big)$$

for $k = 1, \dots, K^{(1)}_m$, and

$$[Y_1, Y_2]^{(k,i)}_{(m)} = \sum_{t_j, t_{j,+} \in \mathcal{G}^{(k,i)} \cap [T_{m-1}, T_m]} \Big(Y_1(t_{j,+}) - Y_1(t_j)\Big)\Big(Y_2(t_{j,+}) - Y_2(t_j)\Big)$$

for $k = 1, 2, \dots, K^{(1)}_m$, $i = 1, \dots, I_m$.

Define

$$\hat{S}^2_{m,j} = N_m^{\frac{1}{3}} \left(\frac{1}{K^{(1)}_{m,j}} \sum_{k=1}^{K^{(1)}_{m,j}} [Y_1, Y_2]^{(k)}_{(m)} - \frac{1}{K^{(2)}_{m,j}} \sum_{i=1}^{I_m} \sum_{k=1}^{K^{(1)}_{m,j}} [Y_1, Y_2]^{(k,i)}_{(m)} \right)^2.$$

The following theorem gives a consistent estimator of ϑ^2 by means of $\hat{S}^2_{m,j}$.

Theorem 2. *Assume (A1)–(A3). If we take $M = M_N \to \infty$ with $M^{1/4}/N_m^{1/3} \to 0$ for each m, then*

$$\widehat{\vartheta^2} := \frac{1}{NT} \sum_{m=1}^{M} \sum_{j=1}^{N} \hat{S}^2_{m,j} = \frac{1}{NT} \sum_{m=1}^{M} \sum_{j=1}^{N} N_m^{\frac{1}{3}} \left(\frac{1}{K^{(1)}_{m,j}} \sum_{k=1}^{K^{(1)}_{m,j}} [Y_1, Y_2]^{(k)}_{(m)} - \frac{1}{K^{(2)}_{m,j}} \sum_{i=1}^{I_m} \sum_{k=1}^{K^{(1)}_{m,j}} [Y_1, Y_2]^{(k,i)}_{(m)} \right)^2 \tag{4.6}$$

is a consistent estimator of ϑ^2.

5. Proofs

Denote $\Delta A(t_j) = A(t_{j+1}) - A(t_j)$ for $A \in \{X_i, Y_i, \epsilon_i, i = 1, 2\}$. In this section, Z_i, $i = 1, 2, 3$, denote standard normal random variables.

Proof of Lemma 1: We observe

$$[Y_1, Y_2]_T = [X_1, X_2]_T + [\epsilon_1, \epsilon_2]_T + [X_1, \epsilon_2]_T + [\epsilon_1, X_2]_T.$$

First, for the asymptotic behavior of $[\epsilon_1, \epsilon_2]_T$, we write

$$\begin{aligned} [\epsilon_1, \epsilon_2]_T &= \sum_{j=0}^{n-1} \Delta\epsilon_1(t_j) \Delta\epsilon_2(t_j) \\ &= 2 \sum_{j=1}^{n-1} \epsilon_1(t_j)\epsilon_2(t_j) + [\epsilon_1(t_0)\epsilon_2(t_0) + \epsilon_1(t_n)\epsilon_2(t_n)] - \sum_{j=0}^{n-1} \epsilon_1(t_{j+1})\epsilon_2(t_j) - \sum_{j=0}^{n-1} \epsilon_1(t_j)\epsilon_2(t_{j+1}) \end{aligned}$$

and

$$\frac{1}{\sqrt{n}}[\epsilon_1,\epsilon_2]_T = \xi_{1,n} + \xi_{2,n} - \xi_{3,n} - \xi_{4,n},$$

where

$$\xi_{1,n} := \frac{2}{\sqrt{n}}\sum_{j=1}^{n-1}\epsilon_1(t_j)\epsilon_2(t_j), \qquad \xi_{2,n} := \frac{1}{\sqrt{n}}\left[\epsilon_1(t_0)\epsilon_2(t_0) + \epsilon_1(t_n)\epsilon_2(t_n)\right]$$
$$\xi_{3,n} := \frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}\epsilon_1(t_{j+1})\epsilon_2(t_j), \quad \xi_{4,n} := \frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}\epsilon_1(t_j)\epsilon_2(t_{j+1}).$$

It is clear that $E[\xi_{2,n}^2] = \mathrm{Var}[\xi_{2,n}] = O(E\epsilon_1^2 E\epsilon_2^2/n)$ and thus $\xi_{2,n} = O_p((E\epsilon_1^2E\epsilon_2^2)^{1/2}/\sqrt{n}) \xrightarrow{\mathrm{p}} 0$. To see the asymptotic behavior of $\xi_{1,n}$, we observe

$$\mathrm{Var}\left(\xi_{1,n}\right) = \frac{4}{n}\sum_{j=1}^{n-1}\mathrm{Var}\left(\epsilon_1(t_j)\epsilon_2(t_j)\right) = \frac{4}{n}\sum_{j=1}^{n-1}E\left(\epsilon_1^2(t_j)\right)E\left(\epsilon_2^2(t_j)\right) \to 4E\epsilon_1^2E\epsilon_2^2.$$

By the central limit theorem of i.i.d. sequences, $\xi_{1,n} \xrightarrow{\mathrm{d}} N(0, 4E\epsilon_1^2E\epsilon_2^2)$. Similarly, it can be shown that $\xi_{3,n}$ and $\xi_{4,n}$ follow asymptotically $N(0, E\epsilon_1^2E\epsilon_2^2)$. Since ϵ_is are independent, so are $\xi_{1,n}$, $\xi_{3,n}$ and $\xi_{4,n}$. Hence we have the normality result:

$$\frac{1}{\sqrt{n}}[\epsilon_1,\epsilon_2]_T \xrightarrow{\mathrm{d}} N\left(0, 6E\epsilon_1^2E\epsilon_2^2\right),$$

and furthermore we have

$$[\epsilon_1,\epsilon_2]_T = \left(6nE\epsilon_1^2E\epsilon_2^2\right)^{\frac{1}{2}} Z_1 + \epsilon_1(t_0)\epsilon_2(t_0) + \epsilon_1(t_n)\epsilon_2(t_n) = \left(6nE\epsilon_1^2E\epsilon_2^2\right)^{\frac{1}{2}} Z_1 + O_p\left(\left(E\epsilon_1^2E\epsilon_2^2\right)^{\frac{1}{2}}\right). \quad (5.1)$$

Secondly, we observe

$$\begin{aligned}[X_1,\epsilon_2]_T &= \sum_{j=0}^{n-1}\Delta X_1(t_j)\Delta\epsilon_2(t_j) = \sum_{j=0}^{n-1}\Delta X_1(t_j)\left(\epsilon_2(t_{j+1}) - \epsilon_2(t_j)\right)\\ &= \sum_{j=1}^{n-1}\left[\Delta X_1(t_{j-1}) - \Delta X_1(t_j)\right]\epsilon_2(t_j) + \Delta X_1(t_{n-1})\epsilon_2(t_n) - \Delta X_1(t_0)\epsilon_2(t_0) =: \zeta_{1,n} + \zeta_{2,n},\end{aligned}$$

where

$$\zeta_{1,n} := \sum_{j=1}^{n-1}\left[\Delta X_1(t_{j-1}) - \Delta X_1(t_j)\right]\epsilon_2(t_j), \qquad \zeta_{2,n} := \Delta X_1(t_{n-1})\epsilon_2(t_n) - \Delta X_1(t_0)\epsilon_2(t_0).$$

It is clear that $\mathrm{Var}(\zeta_{2,n}) = O_p(1/n)$, and thus $\zeta_{2,n} = O_p(1/\sqrt{n})$. Note that $\zeta_{1,n}$ is the sum of a martingale triangular array with increment $[\Delta X_1(t_{j-1}) - \Delta X_1(t_j)]\epsilon_2(t_j)$. By (A2) and by the martingale central limit theorem (Hall and Heyde, 1980, Chapter 3), $\zeta_{1,n}$ is asymptotically normal with mean zero, conditionally on the X_1 process. Now we find the asymptotic (conditional) variance of $\zeta_{1,n}$, i.e., the asymptotic (conditional) variance of $[X_1, \epsilon_2]_T$. Observe

$$\mathrm{Var}\left([X_1, \epsilon_2]_T | X_1\right) = \mathrm{Var}\left(\sum_{j=0}^{n-1} \Delta X_1(t_j)\Delta\epsilon_2(t_j) \middle| X_1\right) = \sum_{j=0}^{n-1}\left(\Delta X_1(t_j)\right)^2 \mathrm{Var}\left(\Delta\epsilon_2(t_j)|X_1\right) = [X_1, X_1]_T 2E\epsilon_2^2.$$

The asymptotic behavior of $[\epsilon_1, X_2]_T$ follows in the same way. Therefore, we have

$$[X_1, \epsilon_2]_T = \left(2E\epsilon_2^2[X_1, X_1]_T\right)^{\frac{1}{2}} Z_2 + O_p\left(\frac{1}{\sqrt{n}}\right), \tag{5.2}$$

$$[\epsilon_1, X_2]_T = \left(2E\epsilon_1^2[X_2, X_2]_T\right)^{\frac{1}{2}} Z_3 + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{5.3}$$

It is clear that $[X_1, \epsilon_2]_T$ and $[\epsilon_1, X_2]_T$ are independent and it can be easily shown that $Cov([\epsilon_1, \epsilon_2]_T, [X_1, \epsilon_2]_T) = 0$. Hence by (5.1)–(5.3), the desired asymptotic result in Lemma 1 is completed. □

Proof of Lemma 2: The proof can be seen in Bibinger (2011b, pp. 20–23) whose Proposition A.1 presented the asymptotic normality of the discrete error of the closest synchronous approximation that equals to the refresh time (Bibinger, 2012, p.2418). Thus we omit the detailed proof. □

Proof of Theorem 1: Since the noises are independent of X-processes, the proof is straightforward from Lemmas 1 and 2. □

Proof of Theorem 2: By (4.3) and (4.4), we have

$$\hat{S}^2_{m,j} = \left(\varsigma_{m,1,j}Z_{m,1,j} - \varsigma_{m,2,j}Z_{m,2,j}\right)^2 + o_p(1), \tag{5.4}$$

where $Z_{m,i,j}$, $i = 1, 2$, are standard normal random variables and

$$\varsigma^2_{m,1,j} = \frac{6}{b_j^2 M}E\epsilon_1^2 E\epsilon_2^2 + b_j M^{\frac{1}{2}}\frac{T}{M}\int_{T_{m-1}}^{T_m}\left[1 + \rho^2(t)\right]\sigma_1^2(t)\sigma_2^2(t)H'(t)dt,$$
$$\varsigma^2_{m,2.j} = \frac{6}{b_j^2 M^2}E\epsilon_1^2 E\epsilon_2^2 + b_j M\frac{T}{M}\int_{T_{m-1}}^{T_m}\left[1 + \rho^2(t)\right]\sigma_1^2(t)\sigma_2^2(t)H'(t)dt$$

with $c_m = b_j M^{1/2}$ and $b_j M$ in (4.4), respectively. Note that in (5.4), $o_p(1) = O_p(1/N_m^{1/3})$ by the proof of Theorem 3.1 of Hwang and Shin (2016). Now we take $M = M_N \to \infty$ with $M^{1/4}/N_m^{1/3} \to 0$ for each m as $N \to \infty$, then

$$\hat{S}^2_{m,j} = \varsigma^2_{m,j}Z^2_{m,j} + o_p(1),$$

where $Z_{m,j}$ are standard normal random variables and $\varsigma^2_{m,j} = \varsigma^2_{m,1,j} + \varsigma^2_{m,2,j} - 2\varsigma_{m,1,j}\varsigma_{m,2,j}$, which tends to $b_j T\int_{T_{m-1}}^{T_m}[1 + \rho^2(t)]\sigma_1^2(t)\sigma_2^2(t)H'(t)dt$. Furthermore, we can express

$$\varsigma^2_{m,j} = b_j f_m\left(1 + O_p\left(M^{-\frac{1}{4}}\right)\right),$$

where $f_m := T\int_{T_{m-1}}^{T_m}[1+\rho^2(t)]\sigma_1^2(t)\sigma_2^2(t)H'(t)dt$ and $O_p(M^{-1/4})$-term comes from $\varsigma_{m,1,j}\varsigma_{m,2,j}$: Indeed, $\varsigma_{m,1,j}\varsigma_{m,2,j}$ can be expressed as $b_j f_m M^{-1/4} + o_p(M^{-1/4})$. Thus, we can express

$$\hat{S}^2_{m,j} = b_j f_m Z^2_{m,j}\left(1 + O_p\left(M^{-\frac{1}{4}}\right)\right).$$

Write

$$\frac{1}{N}\sum_{j=1}^{N}\hat{S}^2_{m,j} = \left[\frac{1}{N}\sum_{j=1}^{N}(b_j-\bar{b})Z^2_{m,j} + \frac{1}{N}\sum_{j=1}^{N}\bar{b}Z^2_{m,j}\right] f_m\left(1 + O_p\left(M^{-\frac{1}{4}}\right)\right), \tag{5.5}$$

where $\bar{b} = (1/N)\sum_{j=1}^{N} b_j$. In the first term of the right-hand side of (5.5), $(1/N)\sum_{j=1}^{N}(b_j-\bar{b})Z^2_{m,j}$ can be easily shown to converges to 0 in probability and in the second term, $(1/N)\sum_{j=1}^{N}\bar{b}Z^2_{m,j} \xrightarrow{\text{p}} \bar{b}E[Z^2_{m,j}] = \bar{b}$. Thus, we have

$$\frac{1}{N}\sum_{j=1}^{N}\hat{S}^2_{m,j} \xrightarrow{\text{p}} f_m\bar{b}\left(1 + O_p\left(M^{-\frac{1}{4}}\right)\right),$$

and also since we choose b_j such that $\bar{b} \to 1$, we have

$$\frac{1}{N}\sum_{j=1}^{N}\hat{S}^2_{m,j} \xrightarrow{\text{p}} f_m\left(1 + O_p\left(M^{-\frac{1}{4}}\right)\right).$$

Hence, we obtain

$$\begin{aligned}\frac{1}{NT}\sum_{m=1}^{M}\sum_{j=1}^{N}\hat{S}^2_{m,j} &\xrightarrow{\text{p}} \sum_{m=1}^{M}\frac{f_m}{T}\left(1 + O_p\left(M^{-\frac{1}{4}}\right)\right)\\ &= \left[\sum_{m=1}^{M}\int_{T_{m-1}}^{T_m}\left[1+\rho^2(t)\right]\sigma_1^2(t)\sigma_2^2(t)H'(t)dt\right]\left(1 + O_p\left(M^{-\frac{1}{4}}\right)\right)\\ &= \vartheta^2\left(1 + O_p\left(M^{-\frac{1}{4}}\right)\right)\end{aligned}$$

which goes to ϑ^2 as $M \to \infty$. Therefore $\widehat{\vartheta^2}$ in (4.6) is a consistent estimator of ϑ^2. □

Acknowledgement

This research was supported by the Gachon University research fund of 2015 (GCU-2015-0036).

References

Aït-Sahalia Y, Fan J, and Xiu D (2010). High-frequency covariance estimates with noisy and asynchronous financial data, *Journal of the American Statistical Association*, **105**, 1504–1517.

Aït-Sahalia Y, Mykland PA, and Zhang L (2005). How often to sample a continuous-time process in the presence of market microstructure noise, *The Review of Financial Studies*, **18**, 351–416.

Bandi FM and Russell JR (2008). Microstructure noise, realized variance, and optimal sampling, *The Review of Economic Studies*, **75**, 339–369.

Barndorff-Nielsen OE and Shephard N (2004). Econometric analysis of realized covariation: high frequency based covariance, regression, and correlation in financial economics, *Econometrica*, **72**, 885–925.

Barndorff-Nielsen OL, Hansen PR, Lunde A, and Shephard N (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading, *Journal of Econometrics*, **162**, 149–169.

Bibinger M (2011a). Efficient covariance estimation for asynchronous noisy high-frequency data, *Scandinavian Journal of Statistics*, **38**, 23–45.

Bibinger M (2011b). *Asymptotics of asynchronicity* (technical report), Retrieved Sep 8, 2016, from: http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2011-033.pdf

Bibinger M (2012). An estimator for the quadratic covariation of asynchronously observed Itô processes with noise: Asymptotic distribution theory, *Stochastic Processes and their Applications*, **122**, 2411–2453.

Dovonon P, Goncalves S, and Meddahi N (2013). Bootstrapping realized multivariate volatility measures, *Journal of Econometrics*, **172**, 49–65.

Griffin JE and Oomen RCA (2011). Covariance measurement in the presence of non-synchronous trading and market microstructure noise, *Journal of Econometrics*, **160**, 58–68.

Hall P and Heyde CC (1980). *Martingale Limit Theory and Its Application*, Academic Press, New York.

Hayashi T and Yoshida N (2005). On covariance estimation of non-synchronously observed diffusion processes, *Bernoulli*, **11**, 359–379.

Hwang E and Shin DW (2016). Two-stage stationary bootstrapping for bivariate average realized volatility matrix under market microstructure noise and asynchronicity, Manuscript submitted.

Mykland PA and Zhang L (2002). *ANOVA for diffusions* (technical report), University of Chicago.

Voev V and Lunde A (2007). Integrated covariance estimation using high-frequency data in the presence of noise, *Journal of Financial Econometrics*, **5**, 68–104.

Zhang L, Mykland PA, 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.

Received July 22, 2016; Revised September 2, 2016; Accepted September 2, 2016