*Electronic Research Archive*

http://www.aimspress.com/journal/era

ERA, 32(2): 799–811.
DOI: 10.3934/era.2024038
Received: 02 October 2023
Revised: 23 December 2023
Accepted: 03 January 2024
Published: 12 January 2024

*Research article*

# Estimation of the quadratic variation of log prices based on the Itô semi-martingale

**Erlin Guo**[1,*] **and Patrick Ling**[2]

[1] School of Mathematics and Statistics, Xuzhou University of Technology, Xuzhou 221018, China
[2] Department of Mathematics, Utah Valley University, Orem, USA

* **Correspondence:** Email: gel@xzit.edu.cn.

**Abstract:** As the availability of high-frequency data becomes more widespread, it has become very popular to model random fluctuations of some econometric variables over time using Itô semi-martingale. An emblematic problem is to estimate the quadratic variation, i.e., the integrated volatility of log prices, using noisy high frequency data with endogenous time and jumps. We propose a methodology that combines the multiple sub-grids and thresholds. First, the sub-sample is used to reduce the effect of the noise. Then, the threshold method is used to get rid of the effect of jumps. Finally, the multiple sub-grids method is used to increase the convergence rate. The asymptotic properties, such as consistency and asymptotic normality, are investigated. Simulation is also included to illustrate the performance of the proposed procedure.

**Keywords:** multiple sub-grids; threshold; endogenous; jumps; semi-martingale

## 1. Introduction

As the availability of high-frequency data becomes more widespread, it has become very popular to model random fluctuations of some econometric variables over time using Itô semi-martingale. Specifically, in financial mathematics, it has become very popular to model log asset prices or interest rates using the stochastic processes $X = (X_t)$:

$$dX_t = b_t dt + \sigma_t dW_t, \qquad \text{for} \qquad t \in [0, 1]. \tag{1.1}$$

for $t \in [0, 1]$ [1]. An emblematic problem in econometrics is how to estimate the quadratic variation (the integrated volatility) of log prices, i.e., $\langle X, X\rangle_t = \int_0^t \sigma_s^2 ds$.

A classical estimator of integrated volatility is the realized volatility c.f. [2], based on the discrete time observations

$$X_{t_i} \qquad \text{for} \qquad 0 = t_0 < t_1 < t_2 < \cdots t_n \le T, \tag{1.2}$$

and the estimator is defined as $[X,X]_t^n = \sum_{t_i \le t}(\Delta X_{t_i})^2$, where $\Delta X_{t_i} = X_{t_i} - X_{t_{i-1}}$ for $i \ge 1$. It is well known that $[X,X]_t^n \xrightarrow{P} \langle X,X\rangle_t$ [3]. However, when it comes to the reality, observed high-frequency data often exhibit complex features and complicated structures due to those issues:

- Jumps;
- Market microstructure noise;
- Endogenous in the price sampling times.

For the first issue, two well-behaved estimators are the multiple-power estimator [4, 5] and the realized threshold quadratic variation [6, 7]. One commonly used assumption is that $X_t$ is a jump-diffusion Itô process:

$$dX_t = dX_t^c + dX_t^d \tag{1.3}$$

for $t \in [0,T]$, where $X^c$and $X^d$ are the continuous and jumps terms, whose forms are given in (2.1) and (2.2) later. Under this setting, the quadratic variation of $X$ becomes

$$[X,X]_t = [X^c,X^c]_t + [X^d,X^d]_t = \int_0^t \sigma_s^2 ds + \sum_{0\le s\le t}(\Delta X_s)^2. \tag{1.4}$$

For the second issue, the model commonly used is the discretely observed process with micro-structure noise:

$$Y_{t_i} = X_{t_i}^c + \varepsilon_{t_i}, \qquad \text{for} \qquad i = 0,1,\cdots,n, \tag{1.5}$$

where $\{\varepsilon_{t_i}, i \ge 0\}$ are i.i.d. random variables, satisfying $E(\varepsilon_{t_i}) = 0$, $E(\varepsilon_{t_i}^2) = \sigma^2$, independent of the process $X_t^c$, and the sampling times $\{t_i, i \ge 0\}$ are independent of $X^c$. For estimating an univariate integrated volatility in the presence of microstructure noise, various estimators have been proposed by researchers, such as two-time scale realized volatility [8], multi-scale realized volatility [9], wavelet realized volatility [10], pre-averaging realized volatility [11], kernel realized volatility [12], and a quasi-maximum likelihood estimator [13]. For estimating a multivariate integrated co-volatility, various methods include a quasi-maximum likelihood estimator based on generalized sampling time [14], the pre-averaging realized volatility [15], realized kernel volatility estimator based on a refresh time scheme [16], and multi-scale realized co-volatility based on previous tick data synchronization [17]. For estimating large integrated volatility matrices, methods consist of universal thresholding [18–21], and adaptive thresholding [22].

For the last issue, the sampling times are irregular or random but (conditionally) independent of the price process. Volatility estimation in some special situations, and in a general situation have been studied [23–25]. A detailed discussion on the issue of possible endogenous effect has been provided in a semi-parametric context [26], and the time endogenous effect on volatility estimation has been investigated in a non-parametric setting [27]. When there were $X^c$, $X^d$ and endogenous time, Li et al. [28] developed a procedure that yields a consistent estimator of the integrated volatility. When there were $X^c$, microstructure noise and endogenous time, Li, Zhang and Zheng [29] considered estimators of the volatility and their asymptotic properties. Li and Guo [30] proposed a new estimator of the integrated volatility in the presence of both market micro-structure noise and jumps when sampling times are endogenous, through averaging every $p$ observations that precede each observation in the

sub-sample $\mathcal{S}$ to remove the effect of $\varepsilon$, and the method of cutting off the "big" part to remove the effect of the jump part. They obtained only an asymptotic rate $n^{1/6-\delta}$ for any $\delta > 0$ due to the local averaging of a single sub-grid being used to reduce the effect of microstructure noise.

We must point out the differences between this paper and [31], although the methods of the two articles seem to be similar. A nonparametric procedure, based on a combination of the preaveraging method and threshold technique, is proposed to estimate the integrated volatility of an Itô semi-martingale in the presence of jumps and microstructure noise. However, we propose a methodology that combines threshold and the multiple sub-grids, to estimate the quadratic variation of an Itô semi-martingale in the presence of endogenous time, jumps, and microstructure noise. First, the sub-sample is used to reduce the effect of the noise. Then, the threshold method is used to get rid of the effect of jumps. Finally, the multiple sub-grids method is used to increase the convergence rate. Thus, the circumstances of the model and the estimated methods are both different.

In this paper, we use the sub-sample to reduce the effect of the noise, while using the multiple sub-grids method to increase the convergence rate. Then, we use the threshold method to get rid of the effect of jumps. we attempt to develop an estimator that converges consistently to the integrated volatility in the presence of jumps, micro-structure noise and time endogenous in a general setting. The asymptotic normality of the proposed estimator is also established.

The remainder of the paper is organized as follows. Some assumptions made by the model and introduction to the methodology are discussed in Section 2. The consistency and asymptotic normality results are given in Section 3. In Section 4, simulation results are presented. Some discussions are given in Section 5 and all the technical proofs are given in the Appendix.

## 2. Preliminaries

### *2.1. Model assumptions*

Let $X = (X_t)$ be the log price of a single asset for continuous time $t \geq 0$, which is defined on a stochastic basis $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$. Then, the model (1.3) is called an Itô semi-martingale if it has the form

$$dX_t^c = b_t dt + \sigma_t dW_t, \tag{2.1}$$

$$dX_t^d = \int_{|x|\leq 1} x(\mu - \nu)(dt, dx) + \int_{|x|>1} x\mu(dt, dx), \tag{2.2}$$

where $b$ and $\sigma$ are locally bounded optional processes, $\mu$ is a jump measure compensated by $\nu$; $\nu(dt, dx)$ has the form $dtF_t(dx)$, where $F_t(dx)$ is a transition measure from $\Omega^{(0)} \times R_+$ endowed with the predictable $\sigma-$field into $R/0$, We define $\beta := \inf\{s : \int_{|x|\leq 1} |x|^s F_t(dx) < \infty\}$, which is called the jump activity index in the literature. If $0 \leq \beta < 1$, we also say that $X^d$ has finite variation.

Actually, instead of observing $X_t$, we observe $Y_t$ due to bid-ask spread bounces, differences in trade sizes, et al., where

$$Y_{t_i} = X_{t_i} + \varepsilon_{t_i}, \qquad \text{for} \qquad i = 0, 1, \cdots, n, \tag{2.3}$$

where $\{\varepsilon_{t_i}, i \geq 0\}$ are i.i.d. random variables, satisfying $E(\varepsilon_{t_i}) = 0$, $E(\varepsilon_{t_i}^2) = \sigma_\varepsilon^2$, and have common fourth moments.

Define the quadratic variation of $X$ as

$$\langle X^c, X^c\rangle_t = \int_0^t \sigma_s^2 ds. \tag{2.4}$$

Here, we aim to develop a new estimator for (2.4) and to investigate some asymptotic properties of the proposed estimator in the presence of jumps, micros-structure noise, and time endogenous.

### *2.2. Methodology*

To estimate the quadratic variation of (2.4), in this section, we give a new estimator $\widehat{\langle X^c, X^c\rangle}_t$. First, we need the notation of $\overline{Y}_{t^k_{i,0}}$ on the $k$-th sub-grid to reduce the effect of the noise. Then, we provide $[\overline{Y}, \overline{Y}]_t^{\mathcal{S}_k}$ to get rid of the effect of jumps on the $k$-th grid. Finally, we use the moving average estimator $\widehat{\langle X^c, X^c\rangle}_t$ based on the multiple sub-grids to obtain the optimal rate $n^{1/4-\delta}$. Now, let us describe the estimator in detail.

Denote $N_t = \max\{i : t_i \le t\}$, we assume that $\max_i \Delta t_i \xrightarrow{P} 0$ is driven by some underlying force, for instance, $n \to \infty$, where $n$ (non-random) measures the sampling frequency over the time interval [0,t]. In constructing the local average, we denote $p$ as the number of observations, $q$ as the size of blocks, and both are non-random numbers just as $n$. Define

$$l := \lfloor \frac{n-p}{q} \rfloor,$$

which satisfies that $lq \le n$, and as $p$ shall be taken as $o(n)$, $lq/n \to 1$ as $n \to \infty$. Moreover, for $k = 0, 1, \cdots, q-1$, we define

$$t^k_{i,j} := t_{iq+p-j+k}, \qquad \text{for} \qquad i = 0, 1, \cdots, \qquad \text{and} \qquad j = 1, 2, \cdots, p-1. \tag{2.5}$$

We consider the time endogeneity on the sub-grid level. The sub-sample $\mathcal{S} = \mathcal{S}_k := \{t_{p+k}, t_{q+p+k}, \cdots, t_{iq+p+k}, \cdots\}$ is constructed by choosing every $q$-th observation starting from the $p+k$-th observation from the complete grid. Then, we define

$$\overline{Y}_{t^k_{i,0}} := \frac{1}{p}\sum_{j=0}^{p-1} Y_{t_{iq+p-j+k}}, \qquad \text{for} \qquad i = 0, 1, 2, \cdots, \qquad \text{and} \qquad k = 0, 1, \cdots, q-1, \tag{2.6}$$

where $t^k_{i,j} = t_{iq+p-j+k}$ and recall that $t^k_{i,0} = t_{iq+p+k}$ denotes the $i$-th observation time on the $k$-th sub-grid.

To get rid of the effect of jumps on the $k$-th grid, the realized volatility of the locally averaged $Y$ process is defined as

$$[\overline{Y}, \overline{Y}]_t^{\mathcal{S}_k} := \sum_{t^k_{i,0} \le t} (\Delta \overline{Y}_{t^k_{i,0}})^2 \mathbf{1}_{\{|\Delta \overline{Y}_{t^k_{i,0}}| \le u^k_{i,0}\}} \tag{2.7}$$

where $\Delta \overline{Y}_{t^k_{i,0}} = \overline{Y}_{t^k_{i,0}} - \overline{Y}_{t^k_{i-1,0}}$ for $i \ge 1$.

After correcting the bias due to noise, the threshold estimator $\widehat{\langle X^c, X^c\rangle}_t$ of $\langle X^c, X^c\rangle_t$ is provided as following:

$$< \widehat{X^c, X^c} >_t = \frac{1}{q}\sum_{k=0}^{q-1} [\overline{Y}, \overline{Y}]_t^{\mathcal{S}_k} - \frac{2L_t}{p}(\hat{\sigma_\varepsilon^2})$$

$$= \frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k\le t}(\Delta\overline{Y}_{t_{i,0}^k})^2\mathbf{1}_{\{|\Delta\overline{Y}_{t_{i,0}^k}|\le u_{i,0}^k\}} - \frac{2L_t}{p}\frac{1}{2n}\sum_{i=1}^{n}(\Delta_i Y)^2\mathbf{1}_{\{|\Delta_i Y|\le u_i\}}, \tag{2.8}$$

where $L_t := \max\{i : t_{i,0}^k \le t\}$, $\hat{\sigma_\varepsilon^2} = \frac{1}{2n}\sum_{i=1}^{n}(\Delta_i Y)^2\mathbf{1}_{\{|\Delta_i Y|\le u_i\}}$ is an estimator of $\sigma_\varepsilon^2$, $u_{i,0}^k$, satisfies

$$u_{i,0}^k/(\Delta_{i,0}^k)^{\varpi_1} \to 0, \qquad u_{i,0}^k/(\Delta_{i,0}^k)^{\varpi_2} \to \infty, \qquad \text{for some} \qquad 0 \le \varpi_1 < \varpi_2 < \frac{1}{4} \tag{2.9}$$

and $u_i$ is similar to $u_{i,0}^k$.

## 3. Results

In this section, the limiting behavior of the estimator will be established. To provide the asymptotic results on multiple sub-grids, the following assumptions are needed.

- (1) There is a filtration $(\mathcal{F}_t)_{t\ge0}$ where $(t_i)_{i\ge1}$ are $(\mathcal{F}_t)$-stopping times. Furthermore, the filtration $(\mathcal{F}_t)$ is generated by finitely many continuous martingales.
- (2) $W_t$, $b_t$ and $\sigma_t^2 \ge c > 0$ are adapted to a filtration $(\mathcal{F}_t)$, integrable and locally bounded, where $c$ is non random;
- (3) $\Delta_n = \max_{1\le i\le n}|t_i - t_{i-1}| = O_p(1/n^{1-\eta})$ for some nonnegative constant $\eta$;
- (4) $L_t/l \xrightarrow{P} \int_0^t r_s ds$ in D[0,1], where $r_s$ is an adapted integrable process;
- (5) the microstructure noise sequence $(\varepsilon_{t_i})_{i\ge0}$ consists of independent random variables with mean 0, variance $\sigma_\varepsilon^2$, and common finite third and forth moments, and is independent of $\mathcal{F}_1$.
- (6) $l\sum_{t_i\le t}(\sum_{j=1}^{q-1}\frac{q-j}{q}\Delta Y_{t_{i-j}}\mathbf{1}_{\{|\Delta Y_{t_{i-j}}|\le u_i\}})^2(\Delta Y_{t_i})^2\mathbf{1}_{\{|\Delta_i Y|\le u_i\}} \xrightarrow{P} \int_0^t w_s\sigma_s^4 ds$ for every $t\in[0,1]$, where $w_s\sigma_s^4$ is integrable, and $u_i$ satisfies (2.9);
- (7) $\frac{1}{q}\sum_{k=0}^{q-1}\sqrt{l}\sum_{t_{i,0}^k\le t}(\Delta\bar{Y}_{t_{i,0}^k})^3\mathbf{1}_{\{|\Delta\bar{Y}_{t_{i,0}^k}|\le u_{i,0}^k\}} \xrightarrow{P} \int_0^t \bar{v}_s\sigma_s^3 ds$ for every $t\in[0,1]$, where $\bar{v}_s^2\sigma_s^4$ is integrable, and $u_{i,0}^k$ satisfies (2.9).

**Remark 1.** *If times are exogenous, Condition (6) can be reduced to a similar assumption in [31]. However, when observation times can be endogenous, the limit is expected to be different.*

**Theorem 1.** *Under the models (2.1)–(2.3) and assumptions (1)–(5), suppose that* $\eta \in [0, 1/9)$*, and* $l \sim C_l n^\alpha$ *and* $p \sim C_p n^\alpha$*, for some* $\max(4\eta, 1/3) < \alpha < (1-\eta)/2$ *and positive constants* $C_l$ *and* $C_p$*, we have*

$$< \widehat{X^c, X^c} >_t \xrightarrow{P} \int_0^t \sigma_s^2 ds. \tag{3.1}$$

**Remark 2.** *In such circumstances, this result does not change the integrated variance of the limit process. The asymptotic mean-squared-error (MSE) is invariable, but must be decomposed differently (see Therem 2).*

**Proof** Thanks to a standard localization procedure, we can use a bounded assumption to replace the local bounded in assumptions, while we also assume that the process $X_t$, itself, and thus the jump

process $X_t^d$, is bounded as well. That is, for all results which need the assumption about volatility and Lévy measure, we may assume further that

$$\max |b_t|, |\sigma_t|, |X_t| \leq C, \qquad \text{for some constant } C > 0 \text{ almost surely.} \tag{3.2}$$

Recall that $Y_t = X_t^c + X_t^d + \varepsilon_t = Z_t + X_t^d$.

We can divide the equation into three parts,

$$\begin{aligned} & < \widehat{X^c, X^c} >_t - \int_0^t \sigma_s^2 ds \\ = & \frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k \leq t}[(\Delta\bar{Z}_{t_{i,0}^k} + \Delta\bar{X}^d_{t_{i,0}^k})^2 \mathbf{1}_{\{|\Delta\bar{Y}_{t_{i,0}^k}| \leq u_{i,0}^k\}} - (\Delta\bar{Z}_{t_{i,0}^k})^2] \\ & -\frac{2L_t}{p}\frac{1}{2n}\sum_{i=1}^{n}[(\Delta_i Z + \Delta_i X^d)^2 \mathbf{1}_{\{|\Delta_i Y| \leq u_i\}} - (\Delta_i Z)^2] \\ & +\frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k \leq t}(\Delta\bar{Z}_{t_{i,0}^k})^2 - \frac{2L_t}{p}\frac{1}{2n}\sum_{i=1}^{n}(\Delta_i Z)^2 - \int_0^t \sigma_s^2 ds \\ = & \xi_{11} + \xi_{12} + \xi_{13}, \end{aligned} \tag{3.3}$$

where

$$\xi_{11} = \frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k \leq t}[(\Delta\bar{Z}_{t_{i,0}^k} + \Delta\bar{X}^d_{t_{i,0}^k})^2 \mathbf{1}_{\{|\Delta\bar{Y}_{t_{i,0}^k}| \leq u_{i,0}^k\}} - (\Delta\bar{Z}_{t_{i,0}^k})^2], \tag{3.4}$$

$$\xi_{12} = \frac{2L_t}{p}\frac{1}{2n}\sum_{i=1}^{n}[(\Delta_i Z + \Delta_i X^d)^2 \mathbf{1}_{\{|\Delta_i Y| \leq u_i\}} - (\Delta_i Z)^2], \tag{3.5}$$

$$\xi_{13} = \frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k \leq t}(\Delta\bar{Z}_{t_{i,0}^k})^2 - \frac{2L_t}{p}\frac{1}{2n}\sum_{i=1}^{n}(\Delta_i Z)^2 - \int_0^t \sigma_s^2 ds. \tag{3.6}$$

(1) For $\xi_{11}$, when $|\Delta\bar{Z}_{t_{i,0}^k}| \geq u_{i,0}^k/2$, for an appropriate constant $C$, we have

$$\xi_{11} \leq C|\Delta\bar{Z}_{t_{i,0}^k}|^{2+m}/(u_{i,0}^k)^m, \tag{3.7}$$

when $|\Delta\bar{Z}_{t_{i,0}^k}| < u_{i,0}^k/2$, we have

$$|\xi_{11}| \leq C|\Delta\bar{Z}_{t_{i,0}^k}|^2|\Delta\bar{X}^d_{t_{i,0}^k}|^r/(u_{i,0}^k)^r, \qquad \text{if} \quad |\Delta\bar{Y}_{t_{i,0}^k}| > u_{i,0}^k \tag{3.8}$$

$$|\xi_{11}| \leq C((|\Delta\bar{X}^d_{t_{i,0}^k}| \wedge u_{i,0}^k)^2 + |\Delta\bar{Z}_{t_{i,0}^k}|(|\Delta\bar{X}^d_{t_{i,0}^k}| \wedge u_{i,0}^k)), \qquad \text{if} \quad |\Delta\bar{Y}_{t_{i,0}^k}| \leq u_{i,0}^k \tag{3.9}$$

where $l$ and $r$ are both any positive numbers which may change at different places. By the assumption of boundedness of the parameters, we repeatedly use Hölder's and Burkholder's inequalities, then

$$E(|\Delta\bar{X}^d_{t_{i,0}^k}|^2) \leq C\Delta_{i,0}^k, \tag{3.10}$$

$$E(|\Delta\bar{Z}_{t_{i,0}^k}|^m) \leq C_m(\Delta_{i,0}^k)^{m/2}, \text{ for } m > 0 \tag{3.11}$$

$$E[(|\Delta\bar{X}^d_{t_{i,0}^k}| \wedge u_{i,0}^k)^2] \leq C\Delta_{i,0}^k(u_{i,0}^k)^{2-\beta} \leq C_s\Delta_{i,0}^k(u_{i,0}^k)^{2-s}, \text{ for } 0 < \beta < s < 2. \tag{3.12}$$

We deduce from above inequalities and estimations

$$\begin{aligned} & E|\xi_{11}| \\ \leq\ & C(\frac{|\Delta\bar{Z}_{t_{i,0}^k}|^{2+m}}{(u_{i,0}^l)^m} + \frac{E|\Delta\bar{Z}_{t_{i,0}^k}|^2 E|\Delta\bar{X}^d_{t_{i,0}^k}|^r}{(u_{i,0}^k)^r} + E(|\Delta\bar{X}^d_{t_{i,0}^k}| \wedge u_{i,0}^k)^2 + E|\Delta\bar{Z}_{t_{i,0}^k}|E(|\Delta\bar{X}^d_{t_{i,0}^k}| \wedge u_{i,0}^k)) \\ \leq\ & C[\frac{(\Delta_{i,0}^k)^{\frac{2+m}{2}}}{(u_{i,0}^k)^m} + \frac{(\Delta_{i,0}^k)^{\frac{2+r}{2}}}{(u_{i,0}^k)^r} + \Delta_{i,0}^k(u_{i,0}^k)^{(2-s)} + \Delta_{i,0}^k(u_{i,0}^k)^{1-s/2}]. \end{aligned} \tag{3.13}$$

Let $m = r = 1$, we have that

$$E|\xi_{11}| \leq C\Delta_{i,0}^k[\frac{(\Delta_{i,0}^k)^{\frac{1}{2}}}{u_{i,0}^k} + (u_{i,0}^k)^{(1-s/2)}]. \tag{3.14}$$

By assumption of $u_{i,0}^k$, we have $\frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k \leq t} E|\xi_{11}| \to 0$ uniformly.

(2) For $\xi_{12}$, similar to $\xi_{11}$, we have $\frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k \leq t} E|\xi_{12}| \to 0$ uniformly.

(3) For $\xi_{13}$, the proof is similar to Theorem 1 of [30] or the result of Theorem 2 in [29], we have $\frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k \leq t} E|\xi_{13}| \to 0$ uniformly.

Combining (1), (2) and (3), we can finish the proof of the theorem. □

We will use the concept of stable convergence in the Central Limit Theorem below. A sequence of random variables (r.v.s) $X_n$ converges stably in law to a r.v. $X$ defined on the appropriate extension of the original probability space, if and only if for any set $A \in \mathcal{F}$ and real number $x$, we have

$$\lim_{n\to\infty} P(X_n \leq x, A) = P(X \leq x, A). \tag{3.15}$$

We shall write it as $X_n \xrightarrow{S} X$. An immediate consequence is that for any $\mathcal{F}-$measurable random variable $\sigma$, we have the joint weak convergence $(X_n, \sigma) \Rightarrow (X, \sigma)$. Hence, it is slightly stronger than convergence in law.

Define

$$A(p,q) := \frac{2}{q}\sum_{j=1}^{p-1}(\frac{p^2}{j^2} - \frac{j}{p}). \tag{3.16}$$

**Theorem 2.** *Under the same assumptions in Theorem 1 and assumptions (6) and (7), then, we have*

$$A(p,q) \sim -n^{4\alpha-2}C_lC_p/3, \tag{3.17}$$

*and stably in law,*

$$\begin{aligned} & l^{1/2}(\frac{1}{q}\sum_{k=0}^{q-1}[\overline{Y},\overline{Y}]_t^{S_k} - \frac{2N_t}{pq}\hat{\sigma_\varepsilon^2} - (1 + A(p,q))\int_0^t \sigma_s^2 ds) \\ \Rightarrow\ & \frac{2}{3}\int_0^t \bar{v}_s\sigma_s dX_s^c + \int_0^t [(4w_s - \frac{4}{9}\bar{v}_s^2)\sigma_s^4 + \frac{8C_l^3}{C_p}r_s(\sigma_\varepsilon^2)^2]^{1/2}dB_s, \end{aligned} \tag{3.18}$$

*where $B_t$ is a standard Brownian motion that is independent of $\mathcal{F}_1$.*

**Remark 3.** *The limiting process of (3.18) depends on the underlying X, the reason is that endogeneity of sampling times is existent. The endogeneity induces a bias term which is nonzero if and only if the limit in* $\frac{1}{q}\sum_{k=0}^{q-1}\sqrt{l}\sum_{t_{i,0}^k\le t}(\Delta\bar{Y}_{t_{i,0}^k})^3\mathbf{1}_{\{|\Delta\bar{Y}_{t_{i,0}^k}|\le u_{i,0}^k\}}$ *is no longer zero. The remaining term is the variance of a normal distribution.*

**Proof** Since the jumps of $X_t$ is a finite variation process when $\beta < 1$, we have the following decomposition:

$$X_t' = X_0 + \int_0^t b_{1s}'ds + \int_0^t \sigma_s dW_s, \qquad X_t'' = X_t - X_t' \tag{3.19}$$

where $b_{1s}' = b_s - \int_{|x|\le 1} sF_s(dx)$, $Z_t' = X_t' + \varepsilon_t$ and $X'' = \sum_{s\le t}\Delta X_s$.

Through the decomposition of (3.18), i.e.,

$$\begin{aligned}
& l^{1/2}(\frac{1}{q}\sum_{k=0}^{q-1}[\bar{Y},\bar{Y}]_t^{\mathcal{S}_k} - \frac{2N_t}{pq}\hat{\sigma_\varepsilon^2} - (1+A(p,q))<X^c,X^c>_t)\\
= \;& l^{1/2}\frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k\le t}[(\Delta\bar{Y}_{t_{i,0}^k})^2\mathbf{1}_{\{|\Delta\bar{Y}_{t_{i,0}^k}|\le u_{i,0}^k\}} - (\Delta\bar{Z}'_{t_{i,0}^k})^2]\\
- \;& l^{1/2}\frac{2N_t}{pq}\frac{1}{2N_t}\sum_{i=1}^{N_t}[(\Delta_i Y)^2\mathbf{1}_{\{|\Delta_i Y|\le u_i\}} - (\Delta_i Z')^2]\\
+ \;& l^{1/2}(\frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k}(\Delta\bar{Z}'_{t_{i,0}^k})^2 - \frac{2N_t}{pq}\frac{1}{2N_t}\sum_{i=1}^{N_t}(\Delta_i Z')^2 - (1+A(p,q))<X^c,X^c>_t),
\end{aligned} \tag{3.20}$$

it suffices to show

$$l^{1/2}\frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k\le t}|(\Delta\bar{Y}_{t_{i,0}^k})^2\mathbf{1}_{\{|\Delta\bar{Y}_{t_{i,0}^k}|\le u_{i,0}^k\}} - (\Delta\bar{Z}'_{t_{i,0}^k})^2| \xrightarrow{P} 0, \tag{3.21}$$

$$l^{1/2}\frac{2N_t}{pq}\frac{1}{2N_t}\sum_{i=1}^{N_t}|(\Delta_i Y)^2\mathbf{1}_{\{|\Delta_i Y|\le u_i\}} - (\Delta_i Z')^2| \xrightarrow{P} 0, \tag{3.22}$$

and

$$\begin{aligned}
& l^{1/2}(\frac{1}{q}\sum_{k=0}^{q-1}\sum_{t_{i,0}^k\le t}(\Delta\bar{Z}'_{t_{i,0}^k})^2 - \frac{2N_t}{pq}\frac{1}{2N_t}\sum_{i=1}^{N_t}(\Delta_i Z')^2 - (1+A(p,q))<X^c,X^c>_t)\\
\Rightarrow\; & \frac{2}{3}\int_0^t \bar{v}_s\sigma_s dX_s^c + \int_0^t[(4w_s - \frac{4}{9}\bar{v}_s^2)\sigma_s^4 + \frac{8C_l^3}{C_p}r_s(\sigma_\varepsilon^2)^2]^{1/2}dB_s.
\end{aligned} \tag{3.23}$$

Similar to Theorem 1, we have the following estimates:

$$E(|\Delta\bar{X}''_{t_{i,0}^k}|) \le C\Delta_{i,0}^k, \tag{3.24}$$

$$E(|\Delta\bar{Z}'_{t^k_{i,0}}|^m) \le C(\Delta^k_{i,0})^{m/2}, \qquad \text{for} \qquad m > 0 \tag{3.25}$$

$$E[|\Delta\bar{X}''_{t^k_{i,0}}| \wedge u^k_{i,0}] \le C(\Delta^k_{i,0})(u^k_{i,0})^{1-s}, \qquad \text{for} \qquad \beta < s < 1. \tag{3.26}$$

By repeated use of Hölder's inequality and the inequality

$$(|x| \wedge u^k_{i,0})^2 \le (u^k_{i,0})^{2-m}(|x| \wedge u^k_{i,0})^m, \qquad \text{for} \qquad 0 < m \le 2, \tag{3.27}$$

we get

$$\begin{aligned} & l^{1/2}E|(\Delta\bar{Y}_{t^k_{i,0}})^2\mathbf{1}_{\{\Delta\bar{Y}_{t^k_{i,0}} \le u^k_{i,0}\}} - (\Delta\bar{Z}'_{t^k_{i,0}})^2| \\ \le\ & Cl^{1/2}[\frac{(\Delta^k_{i,0})^{\frac{2+m}{2}}}{(u^k_{i,0})^m} + \frac{(\Delta^k_{i,0})^{r+1}}{(u^k_{i,0})^r} + (\Delta^k_{i,0})^2(u^k_{i,0})^{2-2s} + (\Delta^k_{i,0})^{3/2}(u^k_{i,0})^{1-s}] \\ \le\ & C(l\Delta^k_{i,0})^{1/2}\Delta^k_{i,0}[(\Delta^k_{i,0})^{\frac{m-1}{2}-m\varpi_2} + (\Delta^k_{i,0})^{r-\frac{1}{2}-r\varpi_2} + (\Delta^k_{i,0})^{2(1-s)\varpi_1+1/2} + (\Delta^k_{i,0})^{(1-s)\varpi_1}]. \end{aligned} \tag{3.28}$$

Let $s \to \beta$, and for some large enough $m$ and $r$, we have

- $m(1/2 - \varpi_2) - 1/2 > 0$,
- $r(1 - \varpi_2) - 1/2 > 0$,
- $2(1 - \beta)\varpi_1 + 1/2 > 0$,
- $(1 - \beta)\varpi_1 > 0$,

because $\varpi_1 > 0$ and $\beta < 1$. Thus, (3.21) is proved.

Similar to (3.21), meanwhile, we combine A.3. in [29] and can get (3.22) and (3.23). □

## 4. Simulation study

In this part, three sample sizes $n = 11,700$, $23,400$ and $46,800$ within $T = 1$ are considered, the log price is drawn from the Ornstein-Uhlenbeck process with drift added by a symmetric stable Lévy process, namely,

$$X_t = \int_0^t cos(s)ds + \int_0^t e^{-2(t-s)}dW_s + X^d_t \tag{4.1}$$

where $W_s$ is a standard Brownian motion, and $X^d_t$ is a symmetric $\beta-$stable Lévy process.

There are several tuning parameters ($n$, $l$, $p$, $q$ and $\beta$) in the proposed estimator that have to be determined. For the sampling frequency $n$, we use the average number of transactions per day for the past, say 30 days as an approximation. In Theorem 1, we notice that $l \sim C_l n^\alpha$ and $p \sim C_p n^\alpha$, so, for $(l, p, q)$, we choose an appropriate $p$. Under the following simulation setting, the standard deviation of the noise is $\sigma_\varepsilon := (\sigma^2_\varepsilon)^{1/2}$=0.0005. We choose $p = 5$, which is found to be good enough to reduce the effect of the micro-structure noise. Since the block size $q$ should be larger than $p$, it is chosen to be 20.

The procedure is repeated 1000 times, and the consistency and asymptotic normality of the estimator are examined. We can get the following observations from the simulation results and QQ-plot.

**Table 1.** Simulation results for $\beta = 0.25$ and $\beta = 0.5$ on three samples.

| n | $\beta = 0.25$ (relative bias, s.e., mse) | $\beta = 0.5$ (relative bias, s.e., mse) |
|---|---|---|
| 11700 | (−0.0766, 0.0530, 0.0087) | (−0.0451, 0.0641, 0.0061) |
| 23400 | (−0.0771, 0.0404, 0.0076) | (−0.0511, 0.0493, 0.0050) |
| 46800 | (−0.0764, 0.0328, 0.0069) | (−0.0514, 0.0414, 0.0044) |

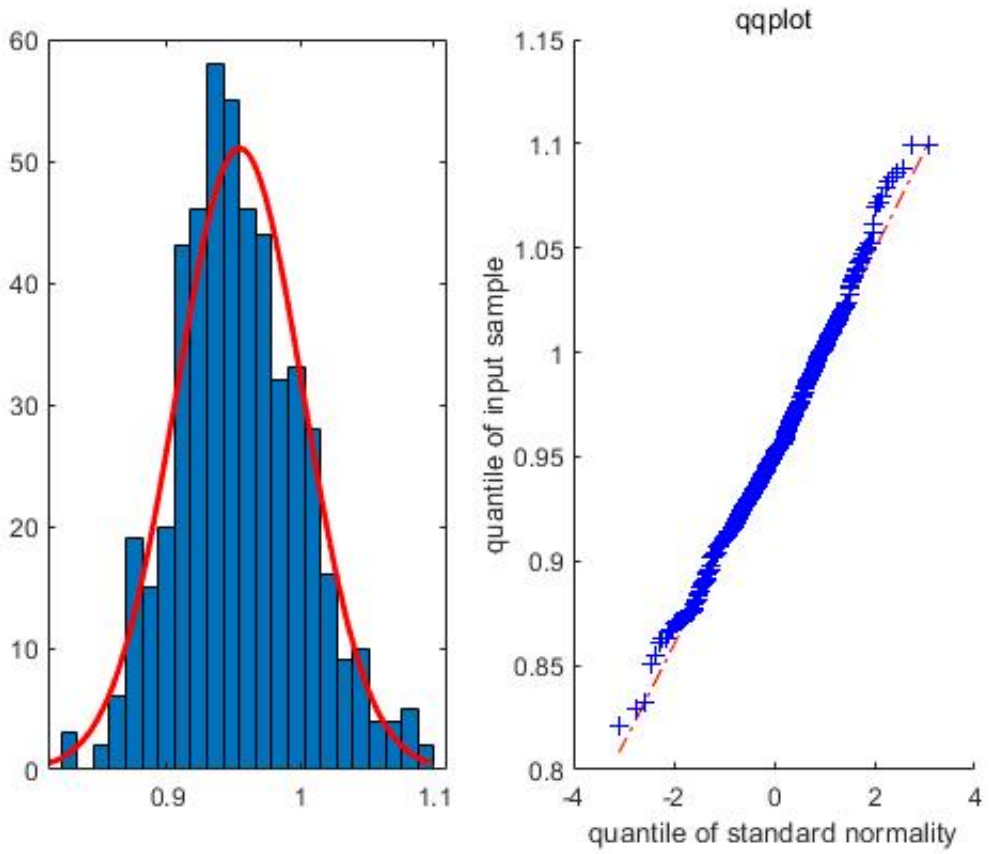


**Figure 1.** QQ-plot for $n = 23400$ and $\beta = 0.5$.

## 5. Conclusions

In this work, based on high-frequency transaction data, we provide a new estimator for the quadratic variation, i.e., integrated volatility, of log prices, in the presence of the endogenous time, micro-structure noise, and jumps. First, we use the sub-sample method to reduce the effect of the noise. Second, we adopt the threshold method to get rid of the effect of jumps. Finally, the multiple sub-grids method is used to increase the rate of convergence. Both the consistency and asymptotic normality of the estimator are investigated. In Theorem 2, if one assumes that $\Delta_n = O_p(1/n)$, then $\eta = 0$, and the convergence rate can be arbitrarily closed to $n^{1/4}$, which is recognized as the optimal convergence rate

in the presence of micro-structure noise. However, with the advance of technology in high-frequency trading, it often involves dozens or even hundreds of assets in financial applications. The corresponding integrated volatility matrix is turned to a high-dimensional problem, which motivates us to develop a new estimator to solve these issues when the observed data have endogenous time, micro-structure noise, jumps, etc.

## Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Acknowledgments

This research was funded by Erlin Guo OF Jiangsu Province grant number BY2022768.

## Conflict of interest

The authors declare no conflicts of interest.

## References

1. J. Q. Fan, A selective overview of nonparametric methods in financial econometrics (with discussion), *Stat. Sci.*, **20** (2005), 317–357.
2. T. Andersen, T. Bollerslev, F. Diebold, P. Labys, Modeling and forcasting realized volatility, *Econometrica*, **71** (2003), 579–625. https://doi.org/10.1111/1468-0262.00418
3. P. Protter, *Stochastic Integration and Differential Equations*, Springer-Verlag, 2004. https://doi.org/10.2307/978-3-540-00313-4
4. O. E. Barndorff-Nielsen, N. Shephard, Power and bipower variation with stochastic volatility and jumps, *J. Financ. Econ.*, **2** (2004), 1–37. https://doi.org/10.1093/jjfinec/nbh001
5. O. E. Barndorff-Nielsen, N. Shephard, Ecnometrics of testing for jumps in financial economics using bipower variation, *J. Financ. Econ.*, **4** (2006), 1–30. https://doi.org/10.1093/jjfinec/nbi022
6. J. Jacod, Asymptotic properties of realized power variations and related functionals of semimartingales, *Stoch. Proc. Appl.*, **118** (2008), 517–559. https://doi.org/10.1016/J.SPA.2007.05.005
7. C. Mancini, Non-parametric threshold estimation for models with stochastic diffusion coefficients and jumps, *Scand. J. Stat.*, **36** (2009), 270–296. https://doi.org/10.1111/j.1467-9469.2008.00622.x
8. L. Zhang, P. A. Mykland, Y. Aït-Sahalia, A tale of two time scales: Determining integrated volatility with noisy high-frequency data, *J. Am. Stat. Assoc.*, **100** (2005), 1394–1411. https://doi.org/10.2307/27590680
9. L. Zhang, Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach, *Bernoulli*, **12** (2006), 1019–1043. https://doi.org/10.2139/ssrn.619682
10. J. Q. Fan, Y. Z. Wang, Multi-scale jump and volatility analysis for high-frequency financial data, *J. Am. Statist. Assoc.*, **102** (2007), 1349–1362. https://doi.org/10.1198/016214507000001067

11. J. Jacod, Y. Li, P. A. Mykland, M. Podolskij, M. Vetter, Microstructure noise in the continuous case: The pre-averaging approach, *Stoch. Proc. Appl.*, **179** (2009), 2249–2276. https://doi.org/10.1016/j.spa.2008.11.0047
12. O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, N. Shephard, Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise, *Econometrica.*, **76** (2008), 1481–1536. https://doi.org/10.3982/ecta6495
13. D. Xiu, Quasi-maximum likelihood estimation of volatility with high frequency data, *J. Econometrics.*, **159** (2010), 235–250. https://doi.org/10.1016/j.jeconom.2010.07.002
14. Y. Aït-Sahalia, J. Q. Fan, D. Xiu, High-frequency covariance estimates with noisy and asynchronous financial data, *J. Amer. Statist. Assoc.*, **105** (2010), 1504–1517. https://doi.org/10.2139/ssrn.1631344
15. K. Christensen, S. Kinnebrock, M. Podolskij, Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data, *J. Econometrics*, **159** (2010), 116–133. https://doi.org/10.1016/j.jeconom.2010.05.001
16. O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, N. Shephard, Multivariate realized kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading, *J. Econometrics.*, **162** (2011), 149–169. https://doi.org/10.1016/j.jeconom.2010.07.009
17. L. Zhang, Estimating covariation: epps effect, microstructure noise, *J. Econometrics.*, **160** (2011), 33–47. https://doi.org/10.1016/j.jeconom.2010.03.012
18. Y. Z. Wang, J. Zou, Vast volatility matrix estimation for high-frequency financial data, *Ann. Statist.*, **38** (2010), 943–978. https://doi.org/10.1214/09-aos730
19. M. Tao, Y. Z. Wang, X. Chen, Fast convergence rates in estimating large volatility matrices using high-frequency financial data, *Economet. Theor.*, **29** (2013), 11–19. https://doi.org/10.2139/ssrn.3786912
20. M. Tao, Y. Z. Wang, H. Zhou, Fast convergence rates in estimating large volatility matrices using high-frequency financial data, *Ann. Statist.*, **41** (2013), 1816–1864. https://doi.org/10.1214/13-aos1128
21. D. Kim, Y. Z. Wang, J. Zou, Asymptotic theory for large volatility matrix estimation based on high-frequency financial data, *Stoch. Proc. Appl.*, **126** (2016), 3527–3577. https://doi.org/10.1016/j.spa.2016.05.004
22. D. Kim, X. B. Kong, C. X. Li, Y. Z. Wang, Adaptive thresholding for large volatility matrix estimation based on high-frequency financial data, *J. Econometrics*, **203** (2018), 69–79. https://doi.org/10.1016/J.JECONOM.2017.09.006
23. M. Fukasawa, Central limit theorem for the realized volatility based on tick time sampling, *Financ. Stoch.*, **14** (2010), 209–233. https://doi.org/10.1007/s00780-008-0087-3
24. M. Fukasawa, M. Rosenbaum, Central limit theorems for realized volatility under hitting times of an irregular grid, *Stoch. Proc. Appl.*, **122** (2012), 3901–3920. https://doi.org/10.1016/j.spa.2012.08.005

25. M. Fukasawa, Realized volatility with stochastic sampling, *Stoch. Proc. Appl.*, **120** (2010), 829–852. https://doi.org/10.1016/j.spa.2010.02.006

26. E. Renault, B. J. Werker, Causality effects in return volatility measures with random times, *J. Econometrics.*, **160** (2011), 272–279. https://doi.org/10.1016/j.jeconom.2010.03.036

27. Y. Li, E. Renault, P. A. Mykland, L. Zhang, X. Zheng, Realized volatility when sampling times are possibly endogenous, *Economet. Theor.*, **30** (2014), 580–605. https://doi.org/10.1017/s0266466613000418

28. C. X. Li, J. Y. Chen, Z. Liu, B. Y. Jing, On integrated volatility of Ito semimartingales when sampling times are endogenous, *Commun. Stat-Theor. M.*, **43** (2014), 5263–5275. https://doi.org/10.1080/03610926.2012.730169

29. Y. Li, Z. Zhang, X. Zheng, Volatility inference in the presence of both endogenous time and microstructure noise, *Stoch. Proc. Appl.*, **123** (2013), 2696–2727. https://doi.org/10.1016/j.spa.2013.04.002

30. C. X. Li, E. L. Guo, Estimation of the integrated volatility using noisy high-frequency data with jumps and endogeneity, *Commun. Stat-Theor. M.*, **47** (2018), 521–531. https://doi.org/10.1080/03610926.2017.1307403

31. B. Y. Jing, Z. Liu, X. B. Kong, On the estimation of integrated volatility with jumps and microstructure noise, *J. Bus. Econ. Stat.*, **32** (2014), 457–467. http://dx.doi.org/10.1080/07350015.2014.906350


AIMS Press