Mathematics

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

AIMS Mathematics, 8(12): 31070–31091.
DOI: 10.3934/math.20231590
Received: 31 August 2023
Revised: 20 October 2023
Accepted: 12 November 2023
Published: 20 November 2023

Research article

Convergence rate for integrated self-weighted volatility by using intraday high-frequency data with noise

Erlin Guo[1], **Cuixia Li**[1,*], **Patrick Ling**[2] **and Fengqin Tang**[3]

[1] School of Mathematics and Statistics, Xuzhou University of Technology, Xuzhou 221018, China
[2] Department of Mathematics, Utah Valley University, Orem, USA
[3] School of Mathematics Sciences, Huaibei Normal University, Huaibei 235000, China

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

Abstract: High-frequency financial data are becoming increasingly available and need to be analyzed under the current circumstances for the market prices of stocks, currencies, risk analysis, portfolio management and other financial instruments. An emblematic challenge in econometrics is estimating the integrated volatility for financial prices, i.e., the quadratic variation of log prices. Following this point, in this paper, we study the estimation of integrated self-weighted volatility, i.e., the generalized style of integrated volatility, by using intraday high-frequency data with noise. In order to reduce the effect of noise, the "pre-averaging" technique is used. Both the law of large numbers and the central limit theorem of the estimator of integrated self-weighted volatility are established in this paper. Meanwhile, a studentized version is also given in order to make some statistical inferences. At the end of this article, the simulation results obtained to evaluate the accuracy of approximating the sampling distributions of the estimator are displayed.

Keywords: nonparametric estimation; self-weighted volatility; high-frequency data; microstructure noise; Itô process
Mathematics Subject Classification: 60G51, 62G35

1. Introduction

1.1. Related studies

With the rapid development of economics, finance and electronic technology, high-frequency financial data are becoming increasingly available and need to be analyzed. In the USA, one-second records are available from the Trade and Quote (TAQ) database, whereas in China, the tick-by-tick stock transactions can be obtained from the private databases of some fund management companies.

Such data are widely used and gamer much attention, particularly for applications related to the market prices of stocks, currencies, risk analysis, portfolio management and other financial instruments [1–4]. However, it is still a big challenge for data analysts to use such data in finance. Most of the studies on using high-frequency data, by far, have focused on the estimation of volatility, which is a key aspect of risk analysis, portfolio management and so on [5–8].

Theoretically, if the latent log price $X = (X_t)$ follows an Itô process

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s, \tag{1.1}$$

where b and σ are locally bounded optional processes and W is a standard Brownian motion. Then, the integrated volatility is the quadratic variation of X, that is, $< X, X >_t = \int_0^t \sigma_s^2 ds$.

The most commonly used estimator of the integrated volatility is the realized volatility, which is based on discrete-time observations [1, 9]. Realized volatility uses intraday high-frequency data to directly measure the volatility in a general semi-martingale model setting. On the one hand, a few works in the literature claim that the diffusion process is unwise. These scholars think that the release of significant "news" in an efficient market will induce discontinuities or jumps in the price process. In this case, the realized volatility will not work as the jumps will be included in the limit. To deal with the jumps, two widely used methods are available, i.e., the bipower method [10] and the threshold estimator [11, 12]. On the other hand, when it comes to the practical side, observed high-frequency financial data are often contaminated by market microstructure noise. Microstructure noise is induced by various frictions in the trading process, such as the asymmetric information of traders, the discreteness of price change, bid-ask spreads and/or rounding errors. Empirical evidence has shown a visible noise accumulation effect at high frequencies, and there has been a number of studies on estimating quantities of interest with prices observed to have microstructure noise. These include the following: the two-time scale method [13, 14]; the multi-scale method [15]; the kernel method [16]; the pre-averaging method [17]; the quasi-maximum likelihood method [18, 19]; auto-covariance and auto-correlation cases including irregular observation times [20]; dependent noise with irregular tick observation times [21]; multiple transaction cases with noise [22–24]. Based on the likelihood of an MA (Moving Average) model, Da and Xiu [25] proposed a simple volatility estimator that assumes that the observed transaction price follows a continuous-time contaminated Itô semi-martingale.

Owing to the rapid development of computer science technology, estimating the integrated volatility matrix of a large number of assets is a challenging problem. The ARVM estimator is proposed based on the use of contaminated high-frequency data under the sparse integrated volatility matrix [26]. The convergence rate of a large volatility matrix estimator using contaminated high-frequency data is given separately in [27] and [28]. By applying principal component analysis to the spatial Kendall's tau matrix instead of the sample covariance matrix, a high-dimensional factor analysis without moment constraints is considered [29]. Since the integrated volatility matrix often has entries with a wide range of variability, an adaptive thresholding estimator and the convergence rate of the estimator are shown under sparse conditions [30]. A non-parametric estimator for a single asset is proposed in [31], and a threshold estimator for multiple assets is considered in [32] for integrated volatility in the presence of both jumps and microstructure noise. Regarding the integrated volatility matrix of a large number of assets, a thresholding estimator has been constructed in the presence of both jumps and microstructure noise, as presented in [33].

Here, based on the observed intraday high-frequency data with microstructure noise, we perform statistical inference on the integrated self-weighted volatility, given as

$$ISV = \int_0^1 f(X_t)\sigma_t^2 dt, \tag{1.2}$$

where X_t is an Itô process, $f : \mathbb{R}^2 \to \mathbb{R}$ is some real Lipschitz function and σ_t^2 is the spot volatility of X. The study of integrated self-weighted volatility ISV is important for both pricing and hedging purposes in financial econometrics. In the current paper, we focus on the latent price asset as a continuous Itô process; the estimation of the jump-diffusion process will be dealt with in future work.

1.2. Contributions and organization

The integrated self-weighted volatility $ISV = \int_0^1 f(X_t)\sigma_t^2 dt$ is a generalization of the integrated volatility $IV = \int_0^1 \sigma_t^2 dt$. The purpose of introducing a general function f is as follows. Theoretically or practically, to make the processes of setting the interest rate or financial asset pricing more realistic or extend the range of the price from $(0, \infty)$ to the whole real line, we transform the interest rate or price process and then analyze the transformed process. Specifically, after taking the logarithm, the price process can yield negative values [34,35]. For example, the convergence rates of a large volatility matrix estimator in the presence of jumps, microstructure noise and asynchronization are given in [33]. Leveraging a variety of factor models, a pre-averaging-based large covariance matrix estimator using high-frequency transaction prices has been constructed [36]. Suppose that one is interested in the integrated volatility of the transformed price processes of X_t. On one hand, the microstructure noise associated with the log price process becomes different when one uses the observed log price and the latent log price, i.e., $\log\left(X(t_i^n) + \varepsilon(t_i^n)\right) - \log\left(X(t_i^n)\right)$. Unfortunately, the microstructure noise associated with the log price process is not centered at zero conditional on X. But this is a necessary condition for consistency of the estimator of integrated volatility for the log price process in many studies [37–39]. Note that the integrated volatility of $\log X_t$ is of the form $\int_0^1 f(X_t)\sigma_t^2 dt$. On the other hand, the self-weighted estimator can potentially balance the correlation between trading price and volume. For these reasons, we chose to estimate ISV with $f(x) = 1/x^2$. So developing inference schemes for ISV provides a more flexible tool.

The rest of the paper is structured as follows. In Section 2, some assumptions of models are given, and the main results, i.e., the law of large numbers and central limit theory, are presented in Section 3. All of the proofs of the theorems are shown in Section 4. Section 5 provides several experimental evaluations and an application to demonstrate the effectiveness of the proposed method. Some conclusions and future research are proposed in Section 6.

2. Assumptions of models

Let $X = (X_t)$ be the true log price at time t, and let it satisfy (1.1). To state our main results, we use the following assumptions.

Assumption 1. *Processes b and σ are adapted to $\{\mathcal{F}_t\}$, i.e., the natural filtration generated by X; moreover, process b is locally bounded; process σ is a càdlàg; for any $t > s > 0$, $W_t - W_s$ is independent of $\mathcal{F}_s$.*

Assumption 2. *Assume that for some constant* $L,\kappa > 0$ *and* $\alpha \in (0,1]$, $f(x)$ *satisfies that*

$$|f(x) - f(y)| \leq L|x-y|^{\alpha}(1 + |x|^{2\kappa} + |y|^{2\kappa}) \tag{2.1}$$

for any $x, y \in \mathbb{R}$.

However, when it comes to the practical side, observed high-frequency financial data are often contaminated by market microstructure noise. We cannot directly observe X_t; instead, we observe Y_t, where

$$Y_t = X_t + \varepsilon_t. \tag{2.2}$$

Let us make the following assumptions about ε.

Assumption 3. *The microstructure noise* ε *is an i.i.d. process given X with*

$$E(\varepsilon_{t_i^n}|Y) = 0 \tag{2.3}$$

$$\sup_{t_i^n \leq 1} E(|\varepsilon_{t_i^n}|^p|Y) < L_p < \infty, \tag{2.4}$$

for any $p > 0$.

To establish the central limit theorem, a standard structural assumption on the volatility process σ is needed.

Assumption 4. *The volatility functions* $\{\sigma_t, t \geq 0\}$ *satisfy the equation*

$$\sigma_t = \sigma_0 + \int_0^t \tilde{b}_s ds + \int_0^t \tilde{\sigma}_s d\tilde{W}_s, \tag{2.5}$$

where $\tilde{b}$ *and* $\tilde{\sigma}$ *are adapted càdlàg processes with* $\tilde{b}$ *being predictable and locally bounded, and* $\tilde{W}$ *is a standard Brownian motion.*

3. Results

Before displaying the main results, we briefly introduce the estimator of ISV. Let us divide the interval $[0,1]$ into m equal sub-intervals, set $K = \lfloor \frac{n}{m} \rfloor$ and denote

$$\tau_r^k \ = \ \frac{r}{m} + \frac{k-1}{n}, \quad r = 0, 1, \cdots, m-1; \quad k = 1, \cdots, K. \tag{3.1}$$

In order to eliminate the effect of microstructure noise, we select l such that $l \to \infty$ and $\frac{l}{n} \to 0$ as $n \to \infty$. For each $i \geq l$, we denote

$$\overline{Y}_{t_i^n} = \frac{1}{l}\sum_{j=1}^{l} Y_{t_{i-j}^n} \ \textit{and}\ \overline{X}_{t_i^n} = \frac{1}{l}\sum_{j=1}^{l} X_{t_{i-j}^n}, \tag{3.2}$$

and then use the following $I\hat{S}V_n$ to estimate ISV:

$$I\hat{S}V_n = \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(Y_{\tau_r^k} - Y_{\tau_{r-1}^k})^2 f(\overline{Y}_{\tau_{r-1}^k}) - \frac{2m}{2n}\sum_{r=1}^{m}\sum_{k=1}^{K}(Y_{\tau_r^k} - Y_{\tau_r^{k-1}})^2 f(\overline{Y}_{\tau_r^{k-1}}), \tag{3.3}$$

where $\hat{\eta^2} = \frac{1}{2n}\sum_{r=1}^{m}\sum_{k=1}^{K}(Y_{\tau_r^k} - Y_{\tau_r^{k-1}})^2$ is the variance estimator of microstructure noise which will be used to reduce the impact of microstructure noise on integrated self-weighted volatility and yield a better estimator.

Theorem 1. *Under Assumptions 1–3, and given*

$$\frac{1}{m^{\alpha/2}} + \frac{1}{l^{\alpha/2}} + \frac{l^{\max(\frac{\alpha}{2},\frac{1-\alpha}{2})}}{n^{\alpha/2}} \to 0, \qquad as \qquad n \to \infty, \tag{3.4}$$

the sequence of random variables $(\frac{1}{m^{\alpha/2}} + \frac{1}{l^{\alpha/2}} + \frac{l^{\max(\frac{\alpha}{2},\frac{1-\alpha}{2})}}{n^{\alpha/2}})^{-1}(I\hat{S}V_n - ISV)$ *is tight.*

To deduce the central limit theorem, the concept of stable convergence is also needed. Let us state it briefly. A sequence of random variables X_n converges stably in law to a random variable X defined on the appropriate extension of the original probability space, written as $X_n \xrightarrow{S} X$, if and only if for any set $A \in \mathcal{F}_1$ and real number x, we have

$$\lim_{n\to\infty} P(X_n \leq x, A) = P(X \leq x, A). \tag{3.5}$$

Hence, stable convergence is slightly stronger than convergence in law.

Theorem 2. *Assume that Assumptions 1–4 hold, and* $\frac{\sqrt{m}}{l^{\alpha/2}} \to 0$ *and* $\frac{\sqrt{m}l^{\max\{\frac{\alpha}{2},\frac{1-\alpha}{2}\}}}{n^{\alpha/2}} \to 0$. *Then, we have*

$$\sqrt{m}(I\hat{S}V_n - ISV) \xrightarrow{S} \sqrt{\frac{7}{6}}\int_0^1 f(X_t)\sigma_t^2 dB_t, \tag{3.6}$$

where B is a standard Brownian motion defined on an extension of the original space, and it is independent of $\mathcal{F}$.

Remark 1. *In Theorem 1, it is worthy of notice that the upper bound is no more than* $\frac{1}{m^{\alpha/2}}$, *where* $\alpha \in (0, 1]$. *However, in Theorem 2, the upper bound is no more than* $\frac{1}{m^{1/2}}$, *which is faster convergence rate than* $\frac{1}{m^{\alpha/2}}$. *The difference between Theorems 1 and 2 is derived from the perspectives of two aspects:*

(1). The estimator we studied in Theorem 2 is derived from the "pre-averaging" technique which essentially tries to "clean" the contaminated data by smoothing first and then applying the usual statistical procedures, while the estimator of Theorem 1 is derived from the two-time-scale technique which applies the usual statistical procedure to the raw contaminated data first, and then corrects the bias caused by the microstructure noise.

(2). An extra bias appears due to the presence of f in the derivation of our asymptotic normality. This is because we need to open an appropriate window with length l to the left of τ_{r-1}^k *in order to smooth away the noise in the contaminated data.*

To perform statistical inference for the suggested estimator, the asymptotic conditional variance $\int_0^1 f^2(X_t)\sigma_t^4 dt$ needs to be estimated. Using Theorem 1, we can give an estimator for the asymptotic conditional variance:

$$\hat{\Gamma}_n = \frac{m}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(Y_{\tau_r^k} - Y_{\tau_{r-1}^k})^4 f^2(\overline{Y}_{\tau_{r-1}^k}) \to 3\int_0^1 f^2(X_s)\sigma_s^4 ds +$$

$$+[2E\varepsilon^4 + 6(E\varepsilon^2)^2]m^2\int_0^1 f^2(X_s)ds + 12m(E\varepsilon^2)\int_0^1 f^2(X_s)\sigma_s^2 ds. \tag{3.7}$$

Remark 2. *The authors of [13] have shown that*

$$\hat{E\varepsilon^2} = \frac{1}{2n}\sum_{r=1}^{m}\sum_{k=1}^{K}(Y_{\tau_r^k} - Y_{\tau_r^{k-1}})^2, \tag{3.8}$$

$$\hat{E\varepsilon^4} = \frac{1}{2n}\sum_{r=1}^{m}\sum_{k=1}^{K}(Y_{\tau_r^k} - Y_{\tau_r^{k-1}})^4 - 3(\hat{E\varepsilon^2})^2. \tag{3.9}$$

Thus, we obtain the consistent estimators as follows:

$$\hat{U}_1 = \frac{1}{mK}\sum_{k=1}^{K}\sum_{r=1}^{m-1} f^2(\overline{Y}_{\tau_{r-1}^k}) \xrightarrow{P} \int_0^1 f^2(X_s)ds, \tag{3.10}$$

$$\begin{aligned}\hat{U}_2 &= \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(Y_{\tau_r^k} - Y_{\tau_{r-1}^k})^2 f^2(\overline{Y}_{\tau_{r-1}^k}) - \frac{2m}{2n}\sum_{r=1}^{m}\sum_{k=1}^{K}(Y_{\tau_r^k} - Y_{\tau_r^{k-1}})^2 f^2(\overline{Y}_{\tau_r^{k-1}}) \\ &\xrightarrow{P} \int_0^1 f^2(X_s)\sigma_s^2 ds. \end{aligned}\tag{3.11}$$

From these equations, and by using a procedure similar to those for Theorems 1 and 2, we obtain a consistent estimator of $\int_0^1 f^2(X_s)\sigma_s^4 ds$*:*

$$\begin{aligned}\hat{C}_n &=: \frac{1}{3}\hat{\Gamma}_n - \frac{m^2}{3}[2\hat{E\varepsilon^4} + 6(\hat{E\varepsilon^2})^2]\hat{U}_1 - 4m\hat{E\varepsilon^2}\hat{U}_2 \\ &\xrightarrow{P} \int_0^1 f^2(X_s)\sigma_s^4 ds. \end{aligned}\tag{3.12}$$

Consequently, a studentized version of the central limit theorem is as follows.

Theorem 3. *Assume that the conditions in Theorem 2 are satisfied; then, we have*

$$\frac{\sqrt{m}(I\hat{S}V_n - ISV)}{\sqrt{\frac{7}{6}\hat{C}_n}} \xrightarrow{S} \mathcal{N}(0,1), \tag{3.13}$$

where $\mathcal{N}(0,1)$ *is a standard normal variable independent of* $\mathcal{F}$*.*

4. Proofs of theorems

Because of the local boundedness of b and σ, we can replace the local boundedness assumptions with boundedness through a standard localization course. As a consequence, process Y has uniformly bounded moments up to any powers on the interval [0,1], i.e., for some constant $C > 0$,

$$\max\{|b_s|, |\sigma_s|, |X_s|\} \leq C. \tag{4.1}$$

Lemma 1. *For any* $\beta > 0$*, we have*

$$E(|\overline{Y}_{\tau_{r-1}^k} - \overline{X}_{\tau_{r-1}^k}|^{2\beta}) \leq C\frac{1}{l^\beta}, \tag{4.2}$$

$$E(|\overline{X}_{\tau^k_{r-1}} - X_{\tau^k_{r-1}}|^{2\beta}) \le C\frac{l^{\max(\beta,1-\beta)}}{n^\beta}, \tag{4.3}$$

$$E(|\overline{Y}_{\tau^k_{r-1}}|^\beta + |\overline{X}_{\tau^k_{r-1}}|^\beta + |X_{\tau^k_{r-1}}|^\beta) \le C, \tag{4.4}$$

$$E(|X_{\tau^k_r} - X_{\tau^k_{r-1}}|^{2\beta}) \le C\frac{1}{m^\beta}. \tag{4.5}$$

Proof. (1) By the definitions of $\overline{Y}_{\tau^k_{r-1}}$ and $\overline{X}_{\tau^k_{r-1}}$, we have, for some given r,

$$\begin{aligned} E|\overline{Y}_{\tau^k_{r-1}} - \overline{X}_{\tau^k_{r-1}}|^{2\beta} &= E|\frac{1}{l}\sum_{j=1}^{l}(Y_{\tau^{k+j}_{r-1}} - X_{\tau^{k+j}_{r-1}})|^{2\beta} \\ &= E|\frac{1}{l}\sum_{j=1}^{l}\varepsilon_{\tau^{k+j}_{r-1}}|^{2\beta} \le C\frac{1}{l^{2\beta}}l^{\beta-1}\sum_{j=1}^{l}E|\varepsilon_{\tau^{k+j}_{r-1}}|^{2\beta} \\ &\le C\frac{1}{l^\beta}. \end{aligned}$$

The last inequality follows from Assumption 3.

(2) We write

$$\begin{aligned} E(|\overline{X}_{\tau^k_{r-1}} - X_{\tau^k_{r-1}}|^{2\beta}) &= E|\frac{1}{l}\sum_{j=1}^{l}(X_{\tau^{k+j}_{r-1}} - X_{\tau^k_{r-1}})|^{2\beta} \\ &\le C\frac{l^{\max(\beta,1-\beta)}}{n^\beta}. \end{aligned}$$

(3) Following from Burkholder-David-Gundy's inequality and the boundedness of b and σ, we can get (4.4) and (4.5). □

Lemma 2. *For any $\beta > 0$, we have*

$$E|f(\overline{Y}_{\tau^k_{r-1}}) - f(\overline{X}_{\tau^k_{r-1}})|^{2\beta} \le C\frac{1}{l^{\alpha\beta}}, \tag{4.6}$$

$$E|f(\overline{X}_{\tau^k_{r-1}}) - f(X_{\tau^k_{r-1}})|^{2\beta} \le C\frac{l^{\max(\alpha\beta,1-\alpha\beta)}}{n^{\alpha\beta}}. \tag{4.7}$$

Proof. Using the Lipschitz property of f, one has

$$\begin{aligned} & E|f(\overline{Y}_{\tau^k_{r-1}}) - f(\overline{X}_{\tau^k_{r-1}})|^{2\beta} \\ \le\ & E[|\overline{Y}_{\tau^k_{r-1}} - \overline{X}_{\tau^k_{r-1}}|^\alpha(1 + |\overline{Y}_{\tau^k_{r-1}}|^{2\kappa} + |\overline{X}_{\tau^k_{r-1}}|^{2\kappa})]^{2\beta} \\ \le\ & C\frac{1}{l^{\alpha\beta}}. \end{aligned}$$

Similar to Lemma 1, we can also get

$$E|f(\overline{X}_{\tau^k_{r-1}}) - f(X_{\tau^k_{r-1}})|^{2\beta} \le C\frac{l^{\max(\alpha\beta,1-\alpha\beta)}}{n^{\alpha\beta}}.$$

Denote that

$$\begin{aligned}
S_1 &= \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(X_{\tau_r^k}-X_{\tau_{r-1}^k})^2 f(X_{\tau_{r-1}^k})-\int_0^1\sigma_s^2 f(X_s)ds,\\
S_2 &= \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(X_{\tau_r^k}-X_{\tau_{r-1}^k})^2 (f(\overline{Y}_{\tau_{r-1}^k})-f(\overline{X}_{\tau_{r-1}^k})),\\
S_3 &= \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(X_{\tau_r^k}-X_{\tau_{r-1}^k})^2 (f(\overline{X}_{\tau_{r-1}^k})-f(X_{\tau_{r-1}^k})),\\
S_4 &= \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(\varepsilon_{\tau_r^k}-\varepsilon_{\tau_{r-1}^k})^2 f(\overline{Y}_{\tau_{r-1}^k})-\frac{m}{n}\sum_{r=1}^{m-1}\sum_{k=1}^{K}(Y_{\tau_r^{k+1}}-Y_{\tau_r^k})^2 f(\overline{Y}_{\tau_r^k}),\\
S_5 &= \frac{2}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(X_{\tau_r^k}-X_{\tau_{r-1}^k})(\varepsilon_{\tau_r^k}-\varepsilon_{\tau_{r-1}^k}) f(\overline{Y}_{\tau_{r-1}^k}).
\end{aligned}$$

□

Lemma 3.

$$(1)\quad E|S_1|\le C\frac{1}{m^{\alpha/2}}. \tag{4.8}$$

$$(2)\quad E|S_2|\le C\frac{1}{l^{\alpha/2}}. \tag{4.9}$$

$$(3)\quad E|S_3|\le C(\frac{l^{\max(\alpha,1-\alpha)}}{n^{\alpha}})^{1/2}. \tag{4.10}$$

$$(4)\quad E|S_4|\le C\frac{m}{n}. \tag{4.11}$$

$$(5)\quad E|S_5|\le C\frac{1}{\sqrt{K}}. \tag{4.12}$$

Proof. (1) For each $k=1,\cdots,K$, we obtain

$$\begin{aligned}
&|\sum_{r=1}^{m-1}(X_{\tau_r^k}-X_{\tau_{r-1}^k})^2 f(X_{\tau_{r-1}^k})-\int_0^1\sigma_s^2 f(X_s)ds|\\
=&\ |\sum_{r=1}^{m-1}[(X_{\tau_r^k}-X_{\tau_{r-1}^k})^2 f(X_{\tau_{r-1}^k})-\int_{\tau_{r-1}^k}^{\tau_r^k}\sigma_s^2 f(X_s)ds]|.
\end{aligned}$$

Thus, by Itô isometry, we have

$$E|S_1| \le E|\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}\int_{\tau_{r-1}^k}^{\tau_r^k}(\sigma_s^2 f(X_{\tau_{r-1}^k})-\sigma_s^2 f(X_s))ds|+C\frac{1}{m},$$

where $C\frac{1}{m}$ is derived from the drift term.

Using the Lipschitz property of f and Hölder's inequality, we have

$$E|S_1|\le\frac{C}{m^{\alpha/2}}.$$

(2) By applying Hölder's inequality, we have

$$\begin{aligned} E|S_2| &\leq \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(E|X_{\tau_r^k} - X_{\tau_{r-1}^k}|^4)^{1/2}(E|f(\overline{Y}_{\tau_{r-1}^k}) - f(\overline{X}_{\tau_{r-1}^k})|^2)^{1/2} \\ &\leq C\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}\frac{1}{m}(\frac{1}{l^\alpha})^{1/2} \\ &\leq C\frac{1}{l^{\alpha/2}}. \end{aligned}$$

(3) By applying Hölder's inequality, we have

$$\begin{aligned} E|S_3| &\leq \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(E|X_{\tau_r^k} - X_{\tau_{r-1}^k}|^4)^{1/2}(E|f(\overline{X}_{\tau_{r-1}^k}) - f(X_{\tau_{r-1}^k})|^2)^{1/2} \\ &\leq C\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}\frac{1}{m}(\frac{l^{\max(\alpha,1-\alpha)}}{n^\alpha})^{1/2} \\ &\leq C(\frac{l^{\max(\alpha,1-\alpha)}}{n^\alpha})^{1/2}. \end{aligned}$$

(4) By [13], it has been proved that $\frac{1}{n}\sum_{r=1}^{m-1}\sum_{k=1}^{K}(Y_{\tau_r^{k+1}} - Y_{\tau_r^k})^2$ is the variance estimator of microstructure noise. So we use $\frac{m}{n}\sum_{r=1}^{m-1}\sum_{k=1}^{K}(Y_{\tau_r^{k+1}} - Y_{\tau_r^k})^2 f(\overline{Y}_{\tau_r^k})$ to reduce the impact of microstructure noise on integrated self-weighted volatility; this yields (4.11).

(5) Since $\varepsilon(\cdot)$ is mutually independent given $\sigma(X)$, we have

$$E(S_5)^2 = \frac{4}{K^2}\sum_{k=1}^{K}\sum_{r=1}^{m-1}E[(X_{\tau_r^k} - X_{\tau_{r-1}^k})(\varepsilon_{\tau_r^k} - \varepsilon_{\tau_{r-1}^k})f(\overline{Y}_{\tau_{r-1}^k})]^2.$$

It follows from Hölder's inequality that

$$E(S_5)^2 \leq \frac{C}{K^2}\sum_{k=1}^{K}\sum_{r=1}^{m-1}[E(X_{\tau_r^k} - X_{\tau_{r-1}^k})^8]^{1/4}E[(\varepsilon_{\tau_r^k} - \varepsilon_{\tau_{r-1}^k})^8]^{1/4}[Ef(\overline{Y}_{\tau_{r-1}^k})^4]^{1/2}.$$

Moreover, because f satisfies the α−Lipschitz Assumption 2, there exists a finite constant C such that $|f(x)| \leq C(1 + |x|^{2\kappa+\alpha})$ for all $x \in \mathbb{R}$. This fact together with the uniform boundedness of any moments of random variables $X(t)$ on $[0, 1]$ yields

$$\sup_{t\in[0,1]} E(f(X(t)))^4 \leq C < \infty.$$

Then, $E(S_5)^2 \leq C\frac{1}{K}$. Hence,

$$E|S_5| \leq C\frac{1}{\sqrt{K}}.$$

□

Proof of Theorem 1. We write

$$\begin{aligned} IS\hat{V}_n - ISV &= \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(X_{\tau_r^k} - X_{\tau_{r-1}^k})^2 f(X_{\tau_{r-1}^k}) - \int_0^1 \sigma_s^2 f(X_s)ds \\ &+\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(X_{\tau_r^k} - X_{\tau_{r-1}^k})^2 (f(\overline{Y}_{\tau_{r-1}^k}) - f(\overline{X}_{\tau_{r-1}^k})) \\ &+\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(X_{\tau_r^k} - X_{\tau_{r-1}^k})^2 (f(\overline{X}_{\tau_{r-1}^k}) - f(X_{\tau_{r-1}^k})) \\ &+\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(\varepsilon_{\tau_r^k} - \varepsilon_{\tau_{r-1}^k})^2 f(\overline{Y}_{\tau_{r-1}^k}) - \frac{m}{n}\sum_{r=1}^{m-1}\sum_{k=1}^{K}(Y_{\tau_r^{k+1}} - Y_{\tau_r^k})^2 f(\overline{Y}_{\tau_r^k}) \\ &+\frac{2}{K}\sum_{k=1}^{K}\sum_{r=1}^{m-1}(X_{\tau_r^k} - X_{\tau_{r-1}^k})(\varepsilon_{\tau_r^k} - \varepsilon_{\tau_{r-1}^k}) f(\overline{Y}_{\tau_{r-1}^k}) \\ &= S_1 + S_2 + S_3 + S_4 + S_5. \end{aligned} \tag{4.13}$$

From Lemma 3, we can get

$$E|IS\hat{V}_n - ISV| \le C(\frac{1}{m^{\alpha/2}} + \frac{1}{l^{\alpha/2}} + \frac{l^{\max(\alpha/2,(1-\alpha)/2)}}{n^{\alpha/2}} + \frac{m}{n} + \frac{1}{K^{1/2}}).$$

Because $\frac{m}{n} = o(\frac{1}{\sqrt{K}})$ and $\frac{1}{\sqrt{K}} = o(\frac{1}{K^{\alpha/2}})$, we always have that

$$E|IS\hat{V}_n - ISV| \le C(\frac{1}{m^{\alpha/2}} + \frac{1}{l^{\alpha/2}} + \frac{l^{\max(\frac{\alpha}{2},\frac{1-\alpha}{2})}}{n^{\alpha/2}}), \tag{4.14}$$

which finishes the proof of Theorem 1. □

Next, let us denote that

$$\begin{aligned} L_1 &= \sum_{r=1}^{[m/2]} E[U_{2r}|\sigma(X_{\tau_{2r-1}^1})] + \sum_{r=1}^{[m/2]} E[U_{2r+1}|\sigma(X_{\tau_{2r}^1})], \\ L_2 &= \sum_{r=1}^{[m/2]}(U_{2r} - E[U_{2r}|\sigma(X_{\tau_{2r-1}^1})]) + \sum_{r=1}^{[m/2]}(U_{2r+1} - E[U_{2r+1}|\sigma(X_{\tau_{2r}^1})]), \end{aligned}$$

where $U_{2r} = \frac{1}{K}\sum_{k=1}^{K}(X_{\tau_{2r}^k} - X_{\tau_{2r-1}^k})^2 f(X_{\tau_{2r-1}^1})$ and $U_{2r+1} = \frac{1}{K}\sum_{k=1}^{K}(X_{\tau_{2r+1}^k} - X_{\tau_{2r}^k})^2 f(X_{\tau_{2r}^1})$. □

Proposition 1.

$$|L_1 - \int_0^1 f(X_s)\sigma)s^2 ds| \le C/m^{\beta}, \ \beta > 1/2. \tag{4.15}$$

Proof. Let $\sigma(X_{\tau_{2r+1}^1})$ and $\sigma(X_{\tau_{2r+2}^1})$ denote the natural filtration respectively generated by $X_{\tau_{2r+1}^1}$ and $X_{\tau_{2r+2}^1}$; then, U_{2r} is $\sigma(X_{\tau_{2r+1}^1})$-adapted and U_{2r+1} is $\sigma(X_{\tau_{2r+2}^1})$-adapted. By the product formula and the definition of σ, for any given $s < t$, we have

$$|E(\sigma_t^2 - \sigma_s^2|\sigma(X_s))| \le C(t - s).$$

Hence,

$$\begin{aligned}\sum_{r=1}^{[m/2]} E[U_{2r}|\sigma(X_{\tau^1_{2r-1}})] &= \sum_{r=1}^{[m/2]} f(X_{\tau^1_{2r-1}})\frac{1}{K}\sum_{k=1}^{K} E(\int_{\tau^k_{2r-1}}^{\tau^k_{2r}} \sigma_s^2 ds|\sigma(X_{\tau^1_{2r-1}}))\\ &= \sum_{r=1}^{[m/2]} f(X_{\tau^1_{2r-1}})\frac{1}{K}\sum_{k=1}^{K}(\sigma^2_{\tau^1_{2r-1}}\frac{1}{m}) + R_{1m},\end{aligned}$$

with $|R_{1m}| \leq \sum_{r=1}^{[m/2]} |f(X_{\tau^1_{2r-1}})|C(\frac{1}{m})^2$, i.e.,

$$\limsup |mR_{1m}| \leq C\int_0^1 |f(X_s)|ds.$$

Similarly, we have

$$\sum_{r=1}^{[m/2]} E[U_{2r+1}|\sigma(X_{\tau^1_{2r}})] = \sum_{r=1}^{[m/2]} f(X_{\tau^1_{2r}})\frac{1}{K}\sum_{k=1}^{K}(\sigma^2_{\tau^1_{2r}}\frac{1}{m}) + R_{2m},$$

with $|R_{2m}| \leq \sum_{r=1}^{[m/2]} |f(X_{\tau^1_{2r}})|C(\frac{1}{m})^2$, and

$$\limsup |mR_{2m}| \leq C\int_0^1 |f(X_s)|ds.$$

Therefore,

$$|L_1 - \int_0^1 f(X_s)\sigma_s^2 ds| \leq C/m^{\beta}, \quad \beta > 1/2.$$

□

Proposition 2.

$$E[(L_2)^2|\sigma(X_{\tau^1_{2r}})] = \frac{7}{3}\frac{1}{m^2}f^2(X_{\tau^1_{2r-1}})\sigma^4_{\tau^1_{2r}} + o_P(\frac{1}{m^2}). \tag{4.16}$$

Proof. Now, let us study the focus of this paper, i.e., L_2.

$$\begin{aligned}L_2 &= \sum_{r=1}^{[m/2]}\{(U_{2r} - E[U_{2r}|\sigma(X_{\tau^1_{2r}})]) + (U_{2r+1} - E[U_{2r+1}|\sigma(X_{\tau^1_{2r}})])\\ &\quad +(E[U_{2r}|\sigma(X_{\tau^1_{2r}})] - E[U_{2r}|\sigma(X_{\tau^1_{2r-1}})])\}.\end{aligned}$$

Up to a smaller order,

$$\begin{aligned}\tilde{L}_2 &= \sum_{r=1}^{[m/2]}\{(U_{2r} - E[U_{2r}|\sigma(X_{\tau^1_{2r}})]) + (U_{2r+1} - E[U_{2r+1}|\sigma(X_{\tau^1_{2r}})])\\ &\quad +(E[U_{2(r+1)}|\sigma(X_{\tau^1_{2r+2}})] - E[U_{2(r+1)}|\sigma(X_{\tau^1_{2r+1}})])\}.\end{aligned}$$

Let $L_{21} = (U_{2r} - E[U_{2r}|\sigma(X_{\tau^1_{2r}})])$, $L_{22} = (U_{2r+1} - E[U_{2r+1}|\sigma(X_{\tau^1_{2r}})])$ and $L_{23} = (E[U_{2(r+1)}|\sigma(X_{\tau^1_{2r+2}})] - E[U_{2(r+1)}|\sigma(X_{\tau^1_{2r+2}})])$. Then,

$$\tilde{L}_2 = \sum_{r=1}^{[m/2]} (L_{21} + L_{22} + L_{23}).$$

Now the summands in the sums of L_{21}, L_{22} and L_{23} are sequences of martingale differences with respect to $\sigma(X_{\tau^1_{2r+2}})$. To us, the remaining work is to find the limit of the conditional variance of $\tilde{L}_2$, i.e., the limit of $\sum_{r=1}^{[m/2]} E[(L_{21} + L_{22} + L_{23})^2|\sigma(X_{\tau^1_{2r}})]$.

(I) By the definition of conditional expectation, we have

$$\begin{aligned} L_{21} &= \frac{1}{K} f(X_{\tau^1_{2r-1}}) \sum_{k=1}^{K} ((X_{\tau^k_{2r}} - X_{\tau^k_{2r-1}})^2 - E[(X_{\tau^k_{2r}} - X_{\tau^k_{2r-1}})^2|\sigma(X_{\tau^1_{2r}})]) \\ &= \frac{1}{K} f(X_{\tau^1_{2r-1}}) \sum_{k=1}^{K} ((X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2 - E[(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2|\sigma(X_{\tau^1_{2r}})]). \end{aligned}$$

So,

$$L_{21}^2 = \frac{1}{K^2} f^2(X_{\tau^1_{2r-1}}) \sum_{k=1}^{K} (L_{21}^k)^2 + \frac{2}{K^2} f^2(X_{\tau^1_{2r-1}}) \sum_{1\le k<j\le K} L_{21}^k L_{21}^j,$$

where,

$$L_{21}^k = (X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2 - E[(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2|\sigma(X_{\tau^1_{2r}})].$$

Using the Itô product formula, we get

$$\begin{aligned} \sum_{k=1}^{K} E[(L_{21}^k)^2|\sigma(X_{\tau^1_{2r}})] &= \sum_{k=1}^{K} (E[(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^4|\sigma(X_{\tau^1_{2r}})] - E^2[(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2|\sigma(X_{\tau^1_{2r}})]) \\ &= 2\sigma^4_{\tau^1_{2r}} \frac{K(K-1)(2K-1)}{6n^2} + o_P(\frac{K^3}{n^2}). \end{aligned}$$

Meanwhile, for $k < j$,

$$\begin{aligned} & E[(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2(X_{\tau^j_{2r}} - X_{\tau^1_{2r}})^2|\sigma(X_{\tau^1_{2r}})] \\ = \; & E[(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^4|\sigma(X_{\tau^1_{2r}})] + E[(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2(X_{\tau^j_{2r}} - X_{\tau^k_{2r}})^2|\sigma(X_{\tau^1_{2r}})] \\ = \; & 3\sigma^4_{\tau^1_{2r}}(\frac{k-1}{n})^2 + \sigma^4_{\tau^1_{2r}} \frac{(k-1)(j-k)}{n^2} + o_P((\frac{k-1}{n})^2 + \frac{(k-1)(j-k)}{n^2}); \end{aligned}$$

then, one gets

$$\sum_{1\le k<j\le K} E[L_{21}^k L_{21}^j|\sigma(X_{\tau^1_{2r}})] = \frac{1}{6}\sigma^4_{\tau^1_{2r}} \frac{K^4}{n^2} + o_P(\frac{K^4}{n^2}).$$

Thus, we obtain that

$$E[L_{21}^2|\sigma(X_{\tau_{2r}^1})] = \frac{1}{3}f^2(X_{\tau_{2r-1}^1})\sigma_{\tau_{2r}^1}^4\frac{1}{m^2} + o_P(\frac{1}{m^2}). \tag{4.17}$$

(II) Denote $L_{22}^k = (X_{\tau_{2r+1}^k} - X_{\tau_{2r}^k})^2 - E[(X_{\tau_{2r+1}^k} - X_{\tau_{2r}^k})^2|\sigma(X_{\tau_{2r}^1})]$; then,

$$E[L_{22}^2|\sigma(X_{\tau_{2r}^1})] = \frac{1}{K^2}f^2(X_{\tau_{2r}^1})(\sum_{k=1}^K E[(L_{22}^k)^2|\sigma(X_{\tau_{2r}^1})] + 2\sum_{1\le k<j\le K} E[L_{22}^k L_{22}^j|\sigma(X_{\tau_{2r}^1})]).$$

Now, for $k < j$, one obtains

$$\begin{aligned} E[L_{22}^k L_{22}^j|\sigma(X_{\tau_{2r}^1})] &= E[(X_{\tau_{2r+1}^k} - X_{\tau_{2r}^j})^2(X_{\tau_{2r+1}^j} - X_{\tau_{2r+1}^k})^2|\sigma(X_{\tau_{2r}^1})] \\ &+ E[(X_{\tau_{2r+1}^k} - X_{\tau_{2r}^j})^4|\sigma(X_{\tau_{2r}^1})] + E[(X_{\tau_{2r}^j} - X_{\tau_{2r}^k})^2(X_{\tau_{2r+1}^j} - X_{\tau_{2r+1}^k})^2|\sigma(X_{\tau_{2r}^1})] \\ &+ E[(X_{\tau_{2r}^j} - X_{\tau_{2r}^k})^2(X_{\tau_{2r+1}^k} - X_{\tau_{2r}^j})^2|\sigma(X_{\tau_{2r}^1})] \\ &- E[(X_{\tau_{2r+1}^k} - X_{\tau_{2r}^k})^2|\sigma(X_{\tau_{2r}^1})]E[(X_{\tau_{2r+1}^j} - X_{\tau_{2r}^j})^2|\sigma(X_{\tau_{2r}^1})] \\ &= \sigma_{\tau_{2r}^1}^4[(\frac{1}{m} + \frac{k-j}{n})(\frac{j-k}{n}) + 3(\frac{1}{m} + \frac{k-j}{n})^2 + (\frac{j-k}{n})^2 \\ &+ (\frac{j-k}{n})(\frac{1}{m} + \frac{k-j}{n}) - \frac{1}{m^2}] + o_P(\frac{1}{m^2}); \end{aligned}$$

thus,

$$E[L_{22}^2|\sigma(X_{\tau_{2r}^1})] = f^2(X_{\tau_{2r}^1})\sigma_{\tau_{2r}^1}^4\frac{1}{m^2} + o_P(\frac{1}{m^2}). \tag{4.18}$$

(III) Let

$$L_{23}^k = E[(X_{\tau_{2r+2}^k} - X_{\tau_{2r+1}^k})^2|\sigma(X_{\tau_{2r+2}^1})] - E[(X_{\tau_{2r+2}^k} - X_{\tau_{2r+1}^k})^2|\sigma(X_{\tau_{2r+1}^1})];$$

then,

$$L_{23} = \frac{1}{K}f(X_{\tau_{2r-1}^1})\sum_{k=1}^K L_{23}^k + \frac{1}{K}(f(X_{\tau_{2r+1}^1}) - f(X_{\tau_{2r-1}^1}))\sum_{k=1}^K L_{23}^k.$$

By the property of the function f, we have

$$\begin{aligned} |\frac{1}{K}(f(X_{\tau_{2r+2}^1}) - f(X_{\tau_{2r-1}^1}))\sum_{k=1}^K L_{23}^k| &\le \frac{1}{K}|X_{\tau_{2r+2}^1} - X_{\tau_{2r-1}^1}|^\alpha(\sum_{k=1}^K|\sigma_{\tau_{2r+2}^1}^2 - \sigma_{\tau_{2r}^1}^2|)\frac{1}{m} \\ &\le \frac{C}{m^{1+\alpha/2}} = o_P(\frac{1}{m^2}). \end{aligned}$$

Now

$$\begin{aligned} L_{23}^k &= (X_{\tau_{2r+2}^1} - X_{\tau_{2r+1}^k})^2 - E[(X_{\tau_{2r+2}^1} - X_{\tau_{2r+1}})^2|\sigma(X_{\tau_{2r+1}^1})] \\ &+ E[(X_{\tau_{2r+2}^k} - X_{\tau_{2r+2}^1})^2|\sigma(X_{\tau_{2r+2}^1})] - E[(X_{\tau_{2r+2}^k} - X_{\tau_{2r+2}^1})^2|\sigma(X_{\tau_{2r+1}^1})] \end{aligned}$$

$$= \quad [(X_{\tau^1_{2r+2}} - X_{\tau^k_{2r+1}})^2 - \sigma^2_{\tau^1_{2r}}(\frac{1}{m} - \frac{k-1}{n})] + (\sigma^2_{\tau^1_{2r+2}} - \sigma^2_{\tau^1_{2r+1}})\frac{k-1}{n} + o_P(\frac{k}{mn});$$

so, we have

$$E[L_{23}^2|\sigma(X_{\tau^1_{2r}})] = \frac{1}{3}f^2(X_{\tau^1_{2r-1}})\sigma^4_{\tau^1_{2r}}\frac{1}{m^2} + o_P(\frac{1}{m^2}). \tag{4.19}$$

(IV) By L_{21}^k and L_{22}^k, we have that

$$E[L_{21}L_{22}|\sigma(X_{\tau^1_{2r}})] = \frac{1}{K^2}f(X_{\tau^1_{2r-1}})f(X_{\tau^1_{2r}})E[(\sum_{k=1}^K L_{21}^k)(\sum_{j=1}^K L_{22}^j)|\sigma(X_{\tau^1_{2r}})].$$

Given

$$\begin{aligned}
& \sum_{k=1}^K\sum_{j=1}^K E[(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2(X_{\tau^j_{2r+1}} - X_{\tau^j_{2r}})^2|\sigma(X_{\tau^1_{2r}})] \\
= & \sum_{1\le k\le j\le K} E[(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2(X_{\tau^j_{2r+1}} - X_{\tau^j_{2r}})^2|\sigma(X_{\tau^1_{2r}})] \\
+ & \sum_{1\le j<k\le K} E[(X_{\tau^k_{2r}} - X_{\tau^j_{2r}})^4|\sigma(X_{\tau^1_{2r}})] \\
+ & \sum_{1\le j<k\le K} E[(X_{\tau^j_{2r}} - X_{\tau^1_{2r}})^2(X_{\tau^k_{2r}} - X_{\tau^j_{2r}})^2|\sigma(X_{\tau^1_{2r}})] \\
+ & \sum_{1\le j<k\le K} E[(X_{\tau^j_{2r}} - X_{\tau^1_{2r}})^2(X_{\tau^j_{2r+1}} - X_{\tau^k_{2r}})^2|\sigma(X_{\tau^1_{2r}})] \\
+ & \sum_{1\le j<k\le K} E[(X_{\tau^k_{2r}} - X_{\tau^j_{2r}})^2(X_{\tau^j_{2r+1}} - X_{\tau^k_{2r}})^2|\sigma(X_{\tau^1_{2r}})] \\
= & \frac{2K^4}{3n^2}\sigma^4_{\tau^1_{2r}} + o_P(\frac{K^4}{n^2})
\end{aligned}$$

and

$$|f(X_{\tau^1_{2r-1}}) - f(X_{\tau^1_{2r}})| \le C|X_{\tau^1_{2r}} - X_{\tau^1_{2r-1}}|^\alpha \le \frac{C}{m^{\alpha/2}},$$

we can obtain

$$2E[L_{21}L_{22}|\sigma(X_{\tau^1_{2r}})] = \frac{1}{3}f^2(X_{\tau^1_{2r}})\sigma^4_{\tau^1_{2r}}\frac{1}{m^2} + o_P(\frac{1}{m^2}). \tag{4.20}$$

(V) By applying the following equation:

$$\begin{aligned}
E[L_{21}L_{23}|\sigma(X_{\tau^1_{2r}})] \quad = & \quad \frac{1}{K^2}f^2(X_{\tau^1_{2r-1}})E[(\sum_{k=1}^K(X_{\tau^k_{2r}} - X_{\tau^1_{2r}})^2 - \sigma^2_{\tau^1_{2r}}\frac{K(K-1)}{2n}) \\
\times & \quad (\sum_{j=1}^K(X_{\tau^1_{2r+2}} - X_{\tau^j_{2r+1}})^2 - \sigma^2_{\tau^1_{2r}}\frac{K(K-1)}{2n})|\sigma(X_{\tau^1_{2r}})] + o_P(\frac{1}{m^2}) = o_P(\frac{1}{m^2}),
\end{aligned}$$

we can get that

$$2E[L_{21}L_{23}|\sigma(X_{\tau^1_{2r}})] = o_P(\frac{1}{m^2}). \tag{4.21}$$

(VI) Similar to (4.21), we have

$$2E[L_{22}L_{23}|\sigma(X_{\tau^1_{2r}})] = \frac{1}{3}f^2(X_{\tau^1_{2r-1}})\sigma^4_{\tau^1_{2r}}\frac{1}{m^2} + o_P(\frac{1}{m^2}). \tag{4.22}$$

Combining (4.17)–(4.22), we have

$$E[(\tilde{L}_2)^2|\sigma(X_{\tau^1_{2r}})] = \frac{7}{3}\frac{1}{m^2}f^2(X_{\tau^1_{2r-1}})\sigma^4_{\tau^1_{2r}} + o_P(\frac{1}{m^2}).$$

□

Proof of Theorem 2. Let

$$\begin{aligned}\hat{I}_n &= \frac{1}{K}\sum_{k=1}^{K}\sum_{r=2}^{m}(X_{\tau^k_r} - X_{\tau^k_{r-1}})^2 f(X_{\tau^k_{r-1}}) = \sum_{r=2}^{m}[\frac{1}{K}\sum_{k=1}^{K}(X_{\tau^k_r} - X_{\tau^k_{r-1}})^2 f(X_{\tau^k_{r-1}})] \\ &:= \sum_{r=2}^{m} U_r := \sum_{r=1}^{[m/2]}(U_{2r} + U_{2r+1}).\end{aligned}$$

By Theorem 1, we have

$$|IS\hat{V}_n - \hat{I}_n| \le C(\frac{1}{l^{\alpha/2}} + \frac{l^{\max(\frac{\alpha}{2},\frac{1-\alpha}{2})}}{n^{\alpha/2}}).$$

So it is enough to prove the result for $\hat{I}_n$. Now, we have

$$\begin{aligned}\hat{I}_n &= \sum_{r=1}^{[m/2]} E[U_{2r}|\sigma(X_{\tau^1_{2r-1}})] + \sum_{r=1}^{[m/2]} E[U_{2r+1}|\sigma(X_{\tau^1_{2r}})] \\ &\quad + \sum_{r=1}^{[m/2]}(U_{2r} - E[U_{2r}|\sigma(X_{\tau^1_{2r-1}})]) + \sum_{r=1}^{[m/2]}(U_{2r+1} - E[U_{2r+1}|\sigma(X_{\tau^1_{2r}})]) \\ &= L_1 + L_2.\end{aligned}$$

Thus, in view of Proposition 1, we have that

$$\sqrt{m}|IS\hat{V}_n - ISV| \le C[\frac{1}{m^{\beta-1/2}} + \frac{\sqrt{m}}{l^{\alpha/2}} + \frac{\sqrt{m}l^{\max(\frac{\alpha}{2},\frac{1-\alpha}{2})}}{n^{\alpha/2}}].$$

Combining Propositions 1 and 2, we have

$$m\sum_{r=1}^{[m/2]} E[(L_{21} + L_{22} + L_{23})^2|\sigma(X_{\tau^1_{2r}})] = \frac{7}{6}\sum_{r=1}^{[m/2]}(f^2(X_{\tau^1_{2r-1}})\sigma^4_{\tau^1_{2r}}\frac{2}{m} + o_P(\frac{1}{m})) \xrightarrow{P} \frac{7}{6}\int_0^1 f^2(X_s)\sigma^4_s ds.$$

□

Proof of Theorem 3. It suffices to prove the consistency of $\hat{C}_n$ with $\int_0^1 f^2(X_t)\sigma^4_t dt$, which is implied by the proof of Theorem 2. □

5. Simulations and applications

5.1. A simulation study

In order to evaluate the performance of the estimator, we consider three sample sizes: $n = 4680$, 7800 and 23400 (which correspond to sampling every 5 seconds, 3 seconds and 1 second) within one trading day ($T = 1$), the latent value is drawn from the geometric Brownian process with drift, namely,

$$X_t = \exp(\int_0^t \mu ds + \int_0^t \sigma dW_s), \tag{5.1}$$

where W_s is a standard Brownian motion. The microstructure noise ε_i's are independent and identically distributed, i.e., $N(0, \omega)$. The variance of microstructure noise was chosen to match the size of the integrated self-weighted volatility. Let t_i be equally spaced in [0,1] with $t_i - t_{i-1} = \frac{1}{n}$, $m = c_m \lfloor n^{\frac{1}{3}} \rfloor$ and $l = \lfloor n^{\frac{1}{3}} \rfloor$ as suggested in Remark 2.

We constructed the observed data according to the definition of $X(t_i)$. The same procedure was repeated 1000 times, and our results, in the form of relative biases, standard errors and mean square errors, are displayed in Table 1. Through the different choices of grid point $c_m \in [2, 4]$ with a fixed step length of 0.2, we calculated the sensitivity of the coverage probabilities(CPs) and confidence intervals(CIs) in Tables 2 and 3. Additionally, the corresponding histogram and QQ-plot are displayed in Figure 1, which verify our central limit theorem. We make the following observations from the simulation results.

(1) In all cases of Table 1, as n increases, all of the biases, standard errors and mean square errors of the estimator tend to decrease; moreover, the estimator is robust against different levels of applied theoretical volatility and variance of microstructure noise. This is in line with our theoretical results.

(2) Tables 2 and 3 show the CPs and CIs at 95% and 90% levels, respectively. Both cases reveal that the CPs and CIs are not sensitive to the choice of c_m.

(3) We demonstrate the central limit theorem in Figure 1 by displaying the histogram and QQ-plot for n=23400, $\sigma^2 = 0.02$, $c_m = 3$.

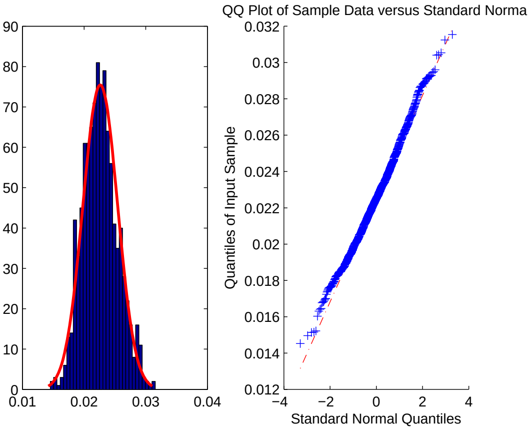

Figure 1. Histogram and QQ-plot results for 1000 values of the estimator, with $\sigma^2 = 0.02$, $c_m = 3$ and $n = 23400$.

Table 1. Estimations of relative bias, standard error(s.e.) and mean square error(MSE).

n	$\sigma^2 = 0.02, \omega = 0.0003^2$	$\sigma^2 = 0.02, \omega = 0.0005^2$
	(relative bias, s.e., MSE)	(relative bias, s.e., MSE)
4680	(0.0011,0.0034,0.0374)	(0.0012,0.0035,0.0262)
11700	(0.0008,0.0030,0.0272)	(0.0007,0.0025,0.0174)
23400	(0.0003,0.0025,0.0136)	(0.0003,0.0013,0.0097)
n	$\sigma^2 = 0.02, \omega = 0.0007^2$	$\sigma^2 = 0.02, \omega = 0.0009^2$
	(relative bias, s.e., MSE)	(relative bias, s.e., MSE)
4680	(0.0012,0.0036,0.0263)	(0.0011,0.0035,0.0411)
11700	(0.0008,0.0029,0.0202)	(0.0010,0.0031,0.0311)
23400	(0.0003,0.0026,0.0099)	(0.0003,0.0025,0.0163)
n	$\sigma^2 = 0.01, \omega = 0.0005^2$	$\sigma^2 = 0.03, \omega = 0.0005^2$
	(relative bias, s.e., MSE)	(relative bias, s.e., MSE)
4680	(0.0009,0.0035,0.0382)	(0.0011,0.0037,0.0367)
11700	(0.0007,0.0022,0.0281)	(0.0010,0.0027,0.0296)
23400	(0.0004,0.0018,0.0086)	(0.0003,0.0021,0.0063)
n	$\sigma^2 = 0.05, \omega = 0.0005^2$	$\sigma^2 = 0.07, \omega = 0.0005^2$
	(relative bias, s.e., MSE)	(relative bias, s.e., MSE)
4680	(0.0016,0.0038,0.0383)	(0.0019,0.0036,0.0376)
11700	(0.0010,0.0030,0.0166)	(0.0007,0.0025,0.0157)
23400	(0.0002,0.0023,0.0102)	(0.0005,0.0009,0.0089)

Table 2. CP and CI results for the asymptotic normal distribution.

	$n = 11700$	$n = 11700$	$n = 23400$	$n = 23400$
c_m	(CP for 95%)	(CI for 95%)	(CP for 95%)	(CI for 95%)
$c_m = 2$	95.7%	(0.0129,0.0259)	95.5%	(0.0135,0.0255)
$c_m = 2.2$	95.6%	(0.0134,0.0271)	94.8%	(0.0152,0.0268)
$c_m = 2.4$	95.4%	(0.0135,0.0254)	94.6%	(0.0150,0.0265)
$c_m = 2.6$	95.3%	(0.0149,0.0271)	94.7%	(0.0154,0.0264)
$c_m = 2.8$	94.6%	(0.0140,0.0252)	94.9%	(0.0153,0.0259)
$c_m = 3$	95.1%	(0.0152,0.0263)	95.2%	(0.0161,0.0241)
$c_m = 3.2$	95.7%	(0.0153,0.0259)	94.6%	(0.0154,0.0248)
$c_m = 3.4$	95.7%	(0.0153,0.0259)	95.7%	(0.0156,0.0250)
$c_m = 3.6$	95.7%	(0.0153,0.0259)	94.9%	(0.0155,0.0245)
$c_m = 3.8$	95.7%	(0.0153,0.0259)	95.5%	(0.0157,0.0239)
$c_m = 4$	95.7%	(0.0153,0.0259)	95.3%	(0.0155,0.0242)

Table 3. CP and CI results for the asymptotic normal distribution.

c_m	$n = 11700$ (CP for 90%)	$n = 11700$ (CI for 95%)	$n = 23400$ (CP for 90%)	$n = 23400$ (CI for 95%)
$c_m = 2$	90.5%	(0.0140,0.0251)	90.3%	(0.0148,0.0245)
$c_m = 2.2$	91.1%	(0.0148,0.0257)	91.0%	(0.0159,0.0261)
$c_m = 2.4$	90.1%	(0.0145,0.0246)	90.8%	(0.0158,0.0256)
$c_m = 2.6$	89.8%	(0.0157,0.0261)	91.1%	(0.0161,0.0256)
$c_m = 2.8$	90.7%	(0.0150,0.0244)	90.1%	(0.0165,0.0249)
$c_m = 3$	89.6%	(0.0161,0.0258)	89.9%	(0.0161,0.0241)
$c_m = 3.2$	90.8%	(0.0152,0.0240)	90.3%	(0.0163,0.0245)
$c_m = 3.4$	90.7%	(0.0165,0.0252)	90.7%	(0.0164,0.0242)
$c_m = 3.6$	90.4%	(0.0157,0.0241)	90.3%	(0.0161,0.0234)
$c_m = 3.8$	90.8%	(0.0165,0.0248)	89.9%	(0.0162,0.0235)
$c_m = 4$	89.9%	(0.0157,0.0238)	90.4%	(0.0164,0.0233)

5.2. An application

In order to investigate the influence of the estimation on the integrated self-weighted volatility in high-frequency data, we collected the intraday transaction prices of 10 stocks from the TAQ database from December 1, 2020 to December 30, 2020. The numbers of observations of the original tick-by-tick data sets for GOOGLE, DELL, JBGS, AAPL, BDN, CACC, EBAY, FF, HMDT and ICE, respectively, were 23,853, 24,478, 20,696, 52,644, 22,701, 23,895, 25,338, 20,265, 18,736 and 24,306. The one-second data sets were constructed by setting the closing price of every one-second interval as the price for the corresponding second. Hence, for each stock, we have approximately second-by-second observations. We constructed an equally spaced data set, because regular spacing is assumed in Theorems 1 and 2. We adopted $c_m = 2$, $m = 60$ and $l = 30$ in the empirical data. Because there is no real integrated volatility for comparison, we calculated the standard deviation of the integrated self-weighted volatility estimators for different periods for 10 stocks in the empirical study. The results are displayed in Table 4.

Table 4. Standard deviation(S. D.) and MSE results for 10 stocks.

	S. D.	MSE
GOOGLE	0.0016	0.0168
DELL	0.0012	0.0019
JBGS	0.0012	0.0008
AAPL	0.0009	0.0028
BDN	0.0004	0.0018
CACC	0.0011	0.0060
EBAY	0.0007	0.0012
FF	0.0005	0.0002
HMDT	0.0010	0.0008
ICE	0.0009	0.0026

6. Conclusions

In this paper, we have proposed a consistent estimator of the integrated self-weighted volatility and demonstrated the central limit theorem of the estimator based on the high-frequency data in the presence of microstructure noise. A studentized version of the proposed estimator has been given. The estimator can potentially be applied to a general semi-martingale and the results can be employed to deal with the statistical inference of volatility.

In our future work, a jump-diffusion model will be considered, where we estimate the integrated self-weighted volatility or integrated self-weighted cross-volatility. Furthermore, we will extend the work to include estimation for multiple-transaction cases and time-endogenous cases under the conditions of more generalized settings than described in this paper.

Use of AI tools declaration

The authors declare they have not used artificial intelligence tools in the creation of this article.

Acknowledgments

This paper was supported by Jiangsu Province (BY2022768), the Qing Lan Project of Jiangsu Province, and NSF China (12071408, 12201235).

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. T. Hendershott, R. Riordan, High frequency trading and price discovery, *J. Economet.*, **148** (2009), 131–148. https://doi.org/10.2139/ssrn.1938769

2. Y. Aït-Sahalia, J. Jacod, Is Brownian motion necessary to model high frequency data? *Ann. Stat.*, **38** (2010), 3093–3128. https://doi.org/10.1214/09-aos749

3. Z. Bai, H. Liu, W. Wong, Enhancement of the applicability of Markowitz's portfolio optimization by utilizing random matrix theory, *Math. Finan.*, **19** (2009), 639–667. https://doi.org/10.1111/j.1467-9965.2009.00383.x

4. J. Liu, F. Longstaff, J. Pan, Dynamic asset allocation with event risk, *J. Financ.*, **58** (2003), 231–259. https://doi.org/10.1111/1540-6261.00523

5. E. Dimson, Risk measurement when shares are subject to infrequent trading, *J. Financ. Econ.*, **7** (1979), 197–226. https://doi.org/10.1016/0304-405X(79)90013-8

6. J. Q. Fan, Y. Y. Li, K. Yu, Vast volatility matrix estimation using high frequency data for portfolio selection, *J. Am. Stat. Assoc.*, **107** (2012), 412–428. https://doi.org/10.1080/1621459.2012.656041

7. Y. Ding, Y. Y. Li, X. H. Zheng, High dimensional minimum variance portfolio estimation under statistical factor models, *J. Economet.*, **222** (2021), 502–515. https://doi.org/10.1016/j.jeconom.2020.07.013

8. T. T. Cai, J. Hu, Y. Y. Li, X. H. Zheng, High-dimensional minimum variance portfolio estimation based on high-frequency data, *J. Economet.*, **214** (2020), 482–494. https://doi.org/10.1016/j.jeconom.2019.04.039

9. O. E. Barndorff-Nielsen, N. Shephard, Econometric analysis of realized volatility and its use in estimating stochastic volatility models, *J. R. Stat. Soc. B.*, **64** (2002), 253–280. https://doi.org/10.1111/1467-9868.00336

10. 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

11. J. Jacod, Asymptotic properties of realized power variation and related functionals of semi-martingales, *Stoch. Proc. Appl.*, **118** (2008), 517–559. https://doi.org/10.1016/j.spa.2007.05.005

12. C. Mancini, Nonparametric threshold estimation for models with stochastic diffusion coefficient and jumps, *Scand. J. Stat.*, **36** (2009), 270–296. https://doi.org/10.1111/j.1467-9469.2008.00622.x

13. L. Zhang, P. 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/10.1198/016214505000000169

14. Y. Aït-Sahalia, P. Mykland, L. Zhang, How often to sample a continuous-time process in the presence of market microstructure noise, *Rev. Financ. Stud.*, **18** (2005), 351–416. https://doi.org/10.1023/A:1004318727672

15. L. Zhang, Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach, *Bernoulli*, **12** (2006), 1019–1043. https://doi.org/10.3150/bj/1165269149

16. O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, N. Shephard, Designing realized kernels to measure ex-post variation of equity prices in the presence of noise, *Econometrica*, **76** (2008), 1481–1536. https://doi.org/10.3982/ECTA6495

17. J. Jacod, Y. Li, P. Mykland, M. Podolskij, M. Vetter, Microstructure noise in the continuous case: the pre-averaging approach, *Stoch. Proc. Appl.*, **119** (2009), 2249–2276. https://doi.org/10.1016/j.spa.2008.11.004

18. D. Xiu, Quasi-maximum likelihood estimation of volatility with high frequency data, *J. Economet.*, **159** (2010), 235–250. https://doi.org/10.1016/j.jeconom.2010.07.002

19. Y. Aït-Sahalia, J. Fan, D. Xiu, High frequency covariance estimates with noisy and asynchronous data, *J. Am. Stat. Assoc.*, **105** (2010), 1504–1517. https://doi.org/10.1198/jasa.2010.tm10163

20. J. Jacod, Y. Li, X. Zheng, Statistical properties of microstructure noise, *Econometrica*, **85** (2017), 1133–1174. https://doi.org/10.3982/ECTA13085

21. J. Jacod, Y. Li, X. Zheng, Estimating the integrated volatility when microstructure noise is dependent and observation times are irregular, *J. Economet.*, **208** (2019), 80–100. https://doi.org/10.2139/ssrn.2659615

22. Z. Liu, Jump-robust estimation of volatility with simultaneous presence of microstructure noise and multiple observations, *Financ. Stoch.*, **21** (2017), 427–469. https://doi.org/10.1007/s00780-017-0325-7

23. Z. Liu, X. Kong, B. Jing, Estimating the integrated volatility using high frequency data with zero durations, *J. Economet.*, **204** (2018), 18–32. https://doi.org/10.1016/j.jeconom.2017.12.008

24. M. Wang, N. Xia, Y. Zhou, On the estimation of high-dimensional integrated covariance matrix based on high-frequency data with multiple transactions, preprint paper, 2021. https://doi.org/10.48550/arXiv.1908.08670

25. R. Da, D. Xiu, When moving-average models meet high-frequency data: uniform inference on volatility, *Econometrica*, **89** (2021), 2787–2825. https://doi.org/10.3982/ECTA15593

26. Y. Z. Wang, J. Zou, Vast volatility matrix estimation for high-frequency financial data, *Ann. Stat.*, **38** (2010), 943–978. https://doi.org/10.1214/09-aos730

27. M. Tao, Y. Z. Wang, H. Zhou, Optimal sparse volatility matrix estimation for high-dimensional Itô process with measurement error, *Ann. Stat.*, **41** (2013), 1816–1864. https://doi.org/10.1214/13-aos1128

28. 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

29. Y. He, X. B. Kong, L. Yu, X. S. Zhang, Large-dimensional factor analysis without moment constraints, *J. Bus. Exon. Stat.*, **40** (2022), 302–312. https://doi.org/10.1080/07350015.2020.1811101

30. 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. Economet.*, **203** (2018), 69–79. https://doi.org/10.1016/J.JECONOM.2017.09.006

31. B. Y. Jing, X. B. Kong, Z. Liu, Modeling high-frequency financial data by pure jump processes, *Ann. Stat.*, **40** (2012), 759–784. https://doi.org/10.1214/12-AOS977

32. B. Y. Jing, C. X. Li, Z. Liu, On estimating the integrated co-volatility using noisy high-frequency data with jumps, *Commun. Stat. Theor. Meth.*, **43** (2013), 3889–3901. https://doi.org/10.1080/03610926.2011.6399746

33. E. L. Guo, C. X. Li, F. Q. Tang, The convergence rates of a large volatility matrix estimator based on noise, jumps, and asynchronization, *Mathematics*, **11** (2023), 1425. https://doi.org/10.3390/math11061425

34. Y. Aït-Sahalia, P. Mykland, L. Zhang, Ultra high frequency volatility estimation with dependent microstructure noise, *J. Economet.*, **160** (2011), 190–203. https://doi,org/10.2139/ssrn.686131

35. K. Christensen, S. Kinnebrock, M. Podolskij, Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data, *J. Economet.*, **72** (2010), 885–925. https://doi.org/10.1016/j.jeconom.2010.05.001

36. C. Dai, K. Lu, D. Xiu, Knowing factors or factor loadings, or neither? Evaluating estimators of large covariance matrices with noisy and asynchronous data, *J. Economet.*, **208** (2019), 43–79. https://doi.org/10.1016/j.jeconom.2018.09.005

37. L. Zhang, Estimating Covariation: Epps effect, microstructure noise, *J. Economet.*, **160** (2010), 33–77. https://doi.org/10.1016/j.jeconom.2010.03.012

38. M. Podolskij, M. Vetter, Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps, *Bernoulli*, **15** (2009), 634–658. https://doi.org/10.17877/DE290R-7733

39. J. Jacod, M. Podolskij, M. Vetter, Limit theorems for moving averages of discretized processes plus noise, *Ann. Stat.*, **38** (2010), 1478–1545. https://doi.org/10.1214/09-AOS756