*Econometric Theory*, **33**, 2017, 331–365.
doi:10.1017/S0266466615000420

# ESTIMATING VOLATILITY FUNCTIONALS WITH MULTIPLE TRANSACTIONS

BING-YI JING
*Hong Kong University of Science and Technology, Hong Kong*
ZHI LIU
*University of Macau and UMacau Zhuhai Research Institute*
XIN-BING KONG
*Soochow University, China*

The phenomenon of multiple transactions at each recording time is a common occurrence for high-frequency financial data because of the heavy trading of the market and limitation of the recording mechanism. This situation has existed for many years, but has become more common in recent years because of heavier trading. Surprisingly, there have been few studies on this important issue, in spite of some ad hoc approaches to treat multiple transactions. In this paper we investigate how to handle multiple transactions, particularly in the context of estimating the integrated volatility and integrated quarticity, which are of great interest in financial econometrics. Two approaches are proposed for this purpose, and their asymptotic properties are investigated. Their performances are confirmed by simulation studies. The estimators are also applied to some real world problems. The work represents only the first step in this direction, and some future research problems are discussed.

## 1. INTRODUCTION

For high-frequency financial data, the phenomenon of multiple transactions (or observations) at the same time is a common occurrence. For example, Table 1 displays the number of transactions of Intel Corporation on August 1, 2000. For the first 18 seconds, there were a total of 245 transactions. The number of transactions per second ranged from 3 to 25, with an average of approximately 14 transactions

The authors would like to thank the Editor and three referees for their very extensive and constructive suggestions that helped to improve this paper considerably. Special thanks go to the Editor, Prof. Peter C. B. Phillips, and the CoEditor, Prof. Eric Renault, for their efforts in polishing the manuscript, which significantly improved the presentation of the paper. We also thank Yingying Li and Xinghua Zheng for their useful discussions. All errors are our own. Jing's research is partially supported by Hong Kong RGC HKUST6019/10P, HKUST6019/12P, and HKUST6022/13P. Liu's research is supported by the FDCT of Macau (No.078/2012/A3 and No.078/2013/A3) and NSFC No.11401607. Kong's research is supported in part by NSFC No.11201080 and in part by Humanity and Social Science Youth Foundation of the Chinese Ministry of Education No. 12YJC910003. Address correspondence to Dr. Zhi Liu at liuzhi@umac.mo.

© Cambridge University Press 2016


https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

**TABLE 1.** Transactions of Intel Corporation in the first 18 seconds on August 1, 2000

| Transaction time | Number of transactions | Number of price changes |
|---|---|---|
| 9:30:02 | 3 | 1 |
| 9:30:03 | 18 | 9 |
| 9:30:04 | 16 | 4 |
| 9:30:05 | 15 | 5 |
| 9:30:06 | 13 | 6 |
| 9:30:07 | 9 | 5 |
| 9:30:08 | 16 | 6 |
| 9:30:09 | 8 | 5 |
| 9:30:10 | 17 | 9 |
| 9:30:11 | 9 | 3 |
| 9:30:12 | 15 | 9 |
| 9:30:13 | 15 | 6 |
| 9:30:14 | 9 | 6 |
| 9:30:15 | 25 | 8 |
| 9:30:16 | 21 | 10 |
| 9:30:17 | 13 | 4 |
| 9:30:18 | 23 | 13 |
| ... | ... | ... |

per second. Among these 245 transactions in the first 18 seconds, there were 105 different transaction prices. Incidentally, multiple transactions occur throughout the trading day, not just at the opening of the market.

The occurrence of multiple transactions is due to the heavy trading of the market and limitation of the recording mechanism. The phenomenon of multiple transactions per unit of time has existed for many years and has become more common in recent years because of heavier trading. For example, Tsay (2005) noted a total of 59,838 transactions (or time intervals) for IBM stock during normal trading hours of 63 trading days from November 1, 1990 to January 31, 1991, of which 6,531 were multiple transactions (i.e., zero time intervals), approximately 10.91% of the total transactions. Nine years later, from December 1 to December 31, 1999, there were a total of 134,120 transactions, among which 22.98% were multiple transactions. Note that the number of intraday tradings has increased sixfold over the span of nine years, which greatly increased the chance of multiple transactions. At the extreme, there were 42 transactions within a given second that occurred twice on December 3, 1999.

Given the common occurrence of multiple transactions, it is natural to ask what effects multiple transactions have on the quantities of interest in financial econometrics and how we should handle them in practice. To our surprise, however, there are few studies on these topics in the literature. The purpose of this paper is to investigate these issues. For simplicity, assume that the log asset price process

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

$X_t = \log S_t$ on the probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ follows

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s, \tag{1}$$

where $W_t$ is a standard Brownian motion, $b_t$ is an adapted and locally bounded process and $\sigma_t$ is a predictable càdlàg process. Our interest here lies in the integrated volatility ($IV_t$):

$$IV_t := [X, X]_t = \int_0^t \sigma_s^2 ds.$$

The observations with multiple transactions can be described as follows. Suppose that there are $M$ distinct transaction times, $0 = s_0 < s_1 < \cdots < s_M = t$, over a fixed time interval $[0, t]$, and that there are $L_i$ ($\geq 1$) transactions recorded at time $s_i$. Then, we can write the multiple observations at each time point as follows:

$$\begin{array}{ll}
\bullet \text{ at time } s_1: & \{X_{s_1,1}, \ldots\ldots, X_{s_1,L_1}\}, \\
\ldots\ldots & \ldots\ldots \\
\bullet \text{ at time } s_m: & \{X_{s_m,1}, \ldots\ldots, X_{s_m,L_m}\}, \\
\ldots\ldots & \ldots\ldots \\
\bullet \text{ at time } s_M: & \{X_{s_M,1}, \ldots\ldots, X_{s_M,L_M}\}.
\end{array} \tag{2}$$

Here we have used the set notation $\mathcal{B}_m := \{X_{s_m,1}, \ldots\ldots, X_{s_m,L_m}\}$ to emphasize that there is no order information for the values in the set $\mathcal{B}_m$. A graphical representation is also given in Figure 1.

We now return to the question of estimating the integrated volatility. In the simple case of no multiple transactions (i.e., $L_m \equiv 1$), the observations become

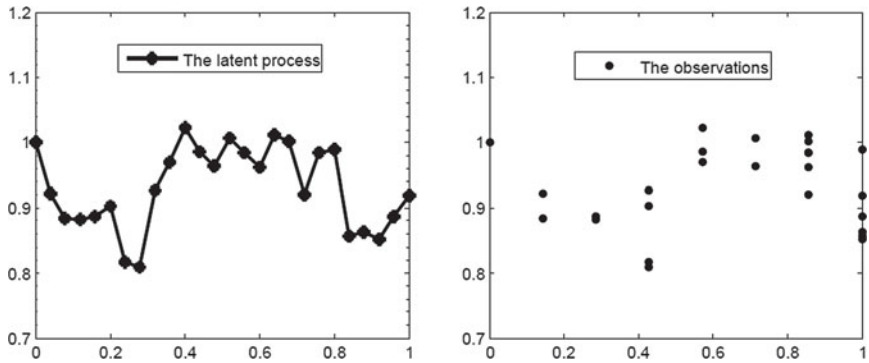



**FIGURE 1.** Left: The latent process Right: The observations

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

$(X_{s_1,1}, ..., X_{s_M,1})$. An obvious consistent estimator of $[X, X]_t$ is the realized volatility $(RV)$:

$$\widehat{[X, X]}_t := \sum_{m=1}^{M} (X_{s_{m,1}} - X_{s_{m-1},1})^2.$$

Clearly, the calculation of the $RV$ for $X$ is based on its increments, where the order information is important. However, in the presence of multiple transactions at a given point in time, the information on their orders may not be available to us, and hence, increments at that time slot cannot be obtained. Furthermore, as noted by Aït-Sahalia, Fan, and Xiu (2010, p. 1507): "in practice, it may happen that the order of consecutive ticks is not recorded correctly." So how do we handle multiple transactions per unit of time in practice?

One commonly used practical approach is to take grouped averages of those multiple transaction prices at each time point $s_m$: $\overline{X}_0 = X_0$ and

$$\overline{X}_{s_m} = \frac{1}{L_m} \sum_{j=1}^{L_m} X_{s_m,j}, \qquad m = 1, ..., M, \tag{3}$$

and use these to approximate their realized integrated volatility:

$$RV_t^{ave} = \sum_{m=1}^{M} \left(\overline{X}_{s_m} - \overline{X}_{s_{m-1}}\right)^2. \tag{4}$$

This is illustrated in Figure 2, where solid circles and squares represent observed and averaged prices, respectively.

However, as the following simple simulation demonstrates, $RV_t^{ave}$ may not be a reasonable estimator of $IV_t = [X, X]_t$ because averaging (smoothing) reduces variability; worse still, it may not even be consistent. In our simple simulation,

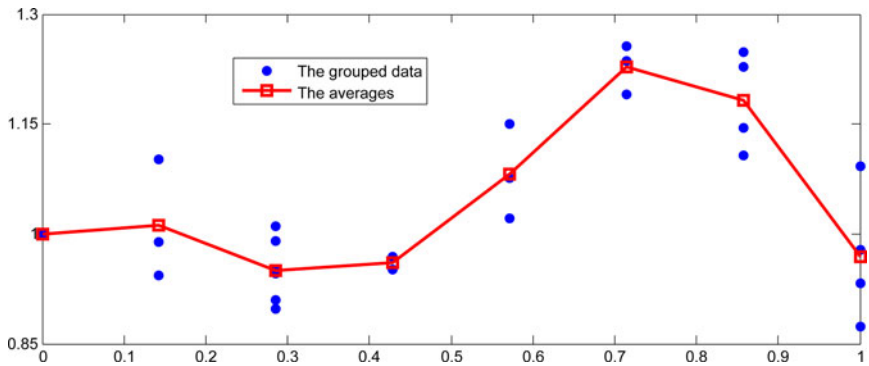



**FIGURE 2.** The explanation of $RV_t^{ave}$.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

we take $X_t = B_t$ for $t \in [0, 1]$, then $IV_1 = 1$. In (2), we take $M = 23,400$ and $L_m \equiv L$, where $L \geq 1$ is a constant to be given later. In other words, in a span of 6.5 trading hours, we record transactions every second (hence $M = 23,400$) with $L$ multiple transactions in each second. The sample can be obtained by first generating $23,400 \times L$ equal-distanced observations from a Brownian motion and then grouping them with equal group size $L$.[1] We then use $RV_t^{ave}$ in (4) to estimate $IV_1 = 1$.

The results for different $L$ are displayed in Table 2 and Figure 3. Clearly, when $L = 1$, i.e., no multiple transactions occur, the $RV_t^{ave}$ estimates the $IV_1 = 1$ almost perfectly. However, as $L$ increases, the performance of the $RV_t^{ave}$ starts to deteriorate sharply for small $L$ and gradually stabilizes for larger $L$. In particular, for $L = 2$, $RV_t^{ave}$ underestimates the $IV_t$ approximately 25%. Overall, $RV_t^{ave}$ performs badly for $L \geq 2$ in the case of Brownian motion.

The poor performance of $RV_t^{ave}$ stimulates interest in finding alternative estimators of $IV_t$ for high-frequency data with multiple observations. From Figure 3, $RV_t^{ave}$ for $L \geq 2$ are all biased downward, and those biases present a clear pattern. Therefore, it might be possible to find alternative estimators that are able to correct those biases. Indeed, one of the proposed methods (see Section 3) does precisely that.

The rest of the paper is arranged as follows. Section 2 gives some basic assumptions. In Sections 3 and 4, we introduce two different estimators of the $IV_t$ in the presence of multiple observations, and establish their limit theorems. We verify the performance of estimators by applying them to some commonly used models in financial literature in Section 5. Applications to real data sets are provided in Section 6. Some open problems and discussions are given in Section 7. The technical proofs are given in Section 8.

## 2. BASIC ASSUMPTIONS

The sample in (2) can be thought of as having been obtained in the following way. Assume that there are transactions $\{X_{t_i}, 1 \leq i \leq n\}$ in $[0, t]$, where $0 = t_0 < t_1 < \cdots < t_n \leq t$. However, due to the limitation of recording mechanisms, we only record transactions at $\{s_m, m \geq 1\}$, where $0 = s_0 < s_1 < s_2 < \cdots < s_M \leq t$. (For example, in our Intel data in Table 1, transactions are recorded every second, so $s_m$ represents the $m$-th second.) Suppose that $L_m$ transactions occurred in $(s_{m-1}, s_m]$ for $m \geq 1$, and their prices were only recorded at time $s_m$. The data set can be

**TABLE 2.** The behavior of realized volatility with different $L$

| $L$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $RV_t^{Ave}$ | 1.0001 | .749 | .705 | .686 | .681 | .677 | .673 | .671 | .669 | .669 | .668 | .667 |

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

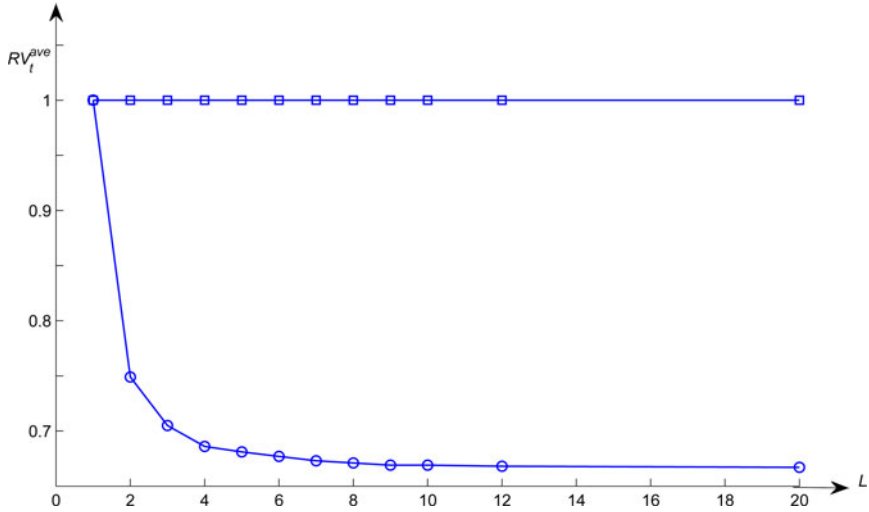



**FIGURE 3.** Pictorial representation of $RV_t^{ave}$ against $L$. The straight line is the target to be estimated.

conveniently represented as follows:

$$\underbrace{X_{t_1},...,X_{t_{N_1}}}_{\text{Observations at } s_1},\ .....,\ \underbrace{X_{t_{N_{m-1}+1}},...,X_{t_{N_m}}}_{\text{Observations at } s_m},.....,\ \underbrace{X_{t_{N_{M-1}+1}},...,X_{t_{N_M}}}_{\text{Observations at } s_M}, \tag{5}$$

where $N_m := \sum_{j=1}^{m} L_j$ denotes the total number of observations in $(0, s_m]$, and $0 = s_0 < t_1 < ... < t_{N_1} \le s_1 < t_{N_1+1} < \cdots < t_{N_2} \le s_2 < \cdots \le s_M = t$; see Figure 1. Denote

$$\Delta_i^n X := X_{t_i} - X_{t_{i-1}}, \qquad \Delta_i^n t := t_i - t_{i-1}, \qquad i = 1, 2, \cdots, n.$$

Assume that the recording times $s_1, ..., s_M$ are equally spaced, and denote $\Delta_M = s_m - s_{m-1}$. This assumption holds for many heavily traded stocks, such as the Intel corporation example given in the Introduction. We further assume that, (unobserved) transaction times in each interval $(s_{m-1}, s_m]$ are equally spaced. As explained in Remark 1, this restriction is not necessary, but it does simplify our exposition. Finally, for a technical reason, we assume that the volatility process $\{\sigma_t, t \ge 0\}$ satisfies

$$\begin{aligned}\sigma_t = \sigma_0 &+ \int_0^t b_s' ds + \int_0^t \sigma_s' dW_s + \int_0^t V_s dB_s \\ &+ \int_0^t \int_E \phi \circ \kappa(s-, x)(\mu - \nu)(ds, dx) \\ &+ \int_0^t \int_E (\kappa - \phi \circ \kappa(s-, x))\mu(dx, ds),\end{aligned} \tag{6}$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

where $B$ is another Brownian motion, independent of $W$, $b'$, $\sigma'$, $V$ are similar to the coefficients of $X$, $\mu$ is a Poisson random measure on $\mathbb{R}\times E$ with predictable compensator $\nu(dt,dz)=dt\lambda(dz)$ and $\lambda$ is a $\sigma$-finite measure on the Polish space $(E,\mathcal{E})$. In addition, $\kappa$ is a predictable function on $\Omega\times\mathbb{R}\times E$, and $\phi$ is a truncation function. This structure of volatility is general enough in the real application. Indeed, it meets many stylized facts of high-frequency data, for example, leverage effect, volatility jumps, and possible common jumps of price and volatility. For a detailed discussion on this structure, see Barndorff-Nielsen, Graversen, Jacod, Podolskij, and Shephard (2006) or Todorov and Tauchen (2010).

In the next two sections, we propose two different estimators of the $IV$ in the presence of multiple transactions.

## 3. APPROACH I: WEIGHTING METHOD

### 3.1. A heuristic argument

It is seen in the Introduction that $RV_t^{ave}=\sum_{m=1}^{M}\left(\overline{X}_{s_m}-\overline{X}_{s_{m-1}}\right)^2$ is biased downward for $IV_t$. We will see that a simple correction will result in a consistent estimator. To motivate our new estimator, let us take a closer look at $RV_t^{ave}$. For instance, when $m=2$, we have

$$\begin{aligned}
\overline{X}_{s_2}-\overline{X}_{s_1} &= \frac{1}{L_2}\sum_{j=1}^{L_2}\left(X_{t_{L_1+j}}-X_{t_1}\right)-\frac{1}{L_1}\sum_{j=1}^{L_1}\left(X_{t_j}-X_{t_1}\right)\\
&= \frac{1}{L_1}\left(\Delta_2^n X+2\Delta_3^n X\ldots+(L_1-1)\Delta_{L_1}^n X\right)\\
&\quad+\frac{1}{L_2}\left(L_2\Delta_{L_1+1}^n X+(L_2-1)\Delta_{L_1+2}^n X\ldots+\Delta_{L_1+L_2}^n X\right)\\
&= \sum_{k=1}^{L_1}\frac{k-1}{L_1}\Delta_k^n X+\sum_{k=1}^{L_2}\left(1-\frac{k-1}{L_2}\right)\Delta_{L_1+k}^n X\\
&= \sum_{k=1}^{L_1+L_2} g_2\left(\frac{k-1}{L_1+L_2}\right)\Delta_k^n X \quad \text{(see Figure 4 (Left panel))},
\end{aligned}$$

where

$$g_2(x)=\begin{cases}\frac{x}{\alpha_2}, & \text{if } x\in(0,\alpha_2)\\ \frac{1-x}{1-\alpha_2}, & \text{if } x\in[\alpha_2,1]\\ 0, & \text{elsewhere.}\end{cases}\qquad \text{with } \alpha_2=\frac{L_1}{L_1+L_2}.$$

See Figure 4 (right panel) for a plot of the weight function $g_2(x)$.

Similarly, for general $m=1,2,\ldots,M$, we have

$$\overline{X}_{s_m}-\overline{X}_{s_{m-1}}=\sum_{k=1}^{L_{m-1}+L_m} g_{m,k}\,\Delta_{N_{m-2}+k}^n X, \tag{7}$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

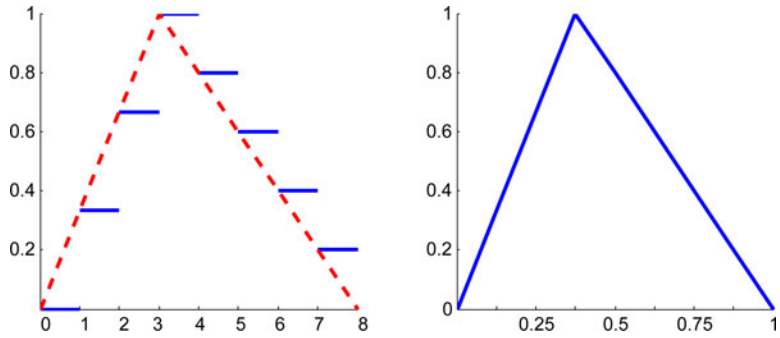



**FIGURE 4.** Left: $g_2\left(\frac{x}{L_1+L_2}\right)$ for $x \in [0, L_1+L_2]$ with $L_1 = 3, L_2 = 5$. Right: Plot of $g_2(x)$

where $g_{m,k} = g_m\left(\frac{k-1}{L_{m-1}+L_m}\right)$ and

$$g_m(x) = \begin{cases} \frac{x}{\alpha_m}, & \text{if } x \in (0, \alpha_m) \\ \frac{1-x}{1-\alpha_m}, & \text{if } x \in [\alpha_m, 1] \\ 0, & \text{elsewhere.} \end{cases} \qquad \text{with } \alpha_m = \frac{L_{m-1}}{L_{m-1}+L_m}.$$

Using the approximation: $\Delta^n_{N_{m-2}+k} X = \int_{t_{N_{m-2}+k-1}}^{t_{N_{m-2}+k}} \sigma_s dW_s \approx \sigma_{s_{m-2}} \Delta^n_{t_{N_{m-2}+k}} W$, we have

$$\begin{aligned} \left(\overline{X}_{s_m} - \overline{X}_{s_{m-1}}\right)^2 &\approx \sigma^2_{s_{m-2}} \left( \sum_{k=1}^{L_{m-1}} g^2_{m,k} \Delta^n_{N_{m-2}+k} t + \sum_{k=L_{m-1}+1}^{L_{m-1}+L_m} g^2_{m,k} \Delta^n_{N_{m-2}+k} t \right) \\ &= \sigma^2_{s_{m-2}} \Delta_M \left( \frac{1}{L_{m-1}} \sum_{k=1}^{L_{m-1}} g^2_{m,k} + \frac{1}{L_m} \sum_{k=L_{m-1}+1}^{L_{m-1}+L_m} g^2_{m,k} \right). \end{aligned} \tag{8}$$

In view of $\sum_{m=1}^{M} \sigma^2_{s_{m-1}} \Delta_M \to \int_0^t \sigma_s^2 ds$, this motivates us to define

$$RV_t^{weight} := \sum_{m=1}^{M} \frac{\left(\overline{X}_{s_m} - \overline{X}_{s_{m-1}}\right)^2}{w_m}, \tag{9}$$

where

$$w_m = \frac{1}{L_{m-1}} \sum_{k=1}^{L_{m-1}} g^2_{m,k} + \frac{1}{L_m} \sum_{k=L_{m-1}+1}^{L_{m-1}+L_m} g^2_{m,k}. \tag{10}$$

The above heuristic argument indicates that $RV_t^{weight}$ should provide a reasonable estimator for $\int_0^t \sigma_s^2 ds$.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

### 3.2. Asymptotic results

Throughout, we use $\xrightarrow{P}$, $\xrightarrow{a.s.}$, $\xrightarrow{\mathcal{L}}$, and $\xrightarrow{\mathcal{S}}$ to denote convergence in probability, almost surely, in distribution and stably in distribution, respectively. For a detailed statement of stable convergence, we refer to Renyi (1963). Let $\mathcal{N}(0,1)$ denote a standard normal random variable (r.v.), and let $m_p = E|\mathcal{N}(0,1)|^p$. We impose an assumption on the number of multiple observations.

**Assumption 1.** $L_m$ are almost surely bounded, i.e., $\sup_m L_m \leq L < \infty$.

THEOREM 1 (Estimating the $IV_t$). *Under Assumption 1, as $M \to \infty$,*

(1). **(Consistency)** *We have*

$$RV_t^{weight} \xrightarrow{P} \int_0^t \sigma_s^2 ds. \tag{11}$$

(2). **(CLT)***: We have*

$$\sqrt{M}\left(RV_t^{weight} - \int_0^t \sigma_s^2 ds\right) \xrightarrow{\mathcal{S}} \sqrt{2}\int_0^t \left(P'(u)\right)^{1/2} \sigma_u^2 dW'_u, \tag{12}$$

*where $W'$ is a Wiener process on an extension of $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ independent of $\mathcal{F}$, and*

$$\hat{P}(u) := M\Delta_M^2 \sum_{s_m \leq u} \frac{(w_m w_{m+1} + \tilde{v}_{m,1} + \tilde{v}_{m,2})}{w_m w_{m+1}} \to P(u), \tag{13}$$

*exists, and is differentiable, and*

$$w_m = \frac{1}{L_{m-1}} \sum_{k=1}^{L_{m-1}-1} \left(\frac{k}{L_{m-1}}\right)^2 + \frac{1}{L_m} \sum_{j=1}^{L_m} \left(1 - \frac{j-1}{L_m}\right)^2,$$

$$\tilde{v}_{m,1} = \frac{1}{L_m^6} \left\{ \sum_{k=1}^{L_m-1} \sum_{j=0}^{k} \left(k^2(L_m - j)^2 + (L_m - k)^2 j^2 + 4kj(L_m - k)(L_m - j)\right) \right\},$$

$$\tilde{v}_{m,2} = -\frac{1}{L_m^6} \sum_{k=1}^{L_m-1} \sum_{j=0}^{L_m-1} (L_m - j)^2 k^2.$$

**Remark 1.** A few words about (13) are in order.

- (13) holds true if $L_m \equiv L$, where it is easily seen that $P(u) = Au$, and hence $P'(u) = A$ with

$$A = 1 + \frac{1}{L^6 w^2} \left\{ \sum_{k=1}^{L-1} \sum_{j=0}^{k} \left((L-k)^2 j^2 + k^2 (L-j)^2 + 4kj(L-k)(L-j)\right) \right\}$$
$$- \frac{1}{L^6 w^2} \left( \sum_{k=1}^{L-1} k^2 \sum_{j=0}^{L-1} (L-j)^2 \right). \tag{14}$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

In particular, when $L_m \equiv 1$, we have $A = 1$, it is the classical central limit theorem of quadratic variation of continuous semi-martingale.

- In reality, the number of arrivals $\{L_m, 1 \le m \le M\}$ is typically different. We now assume that the transaction numbers $L_m$ follow a Poisson distribution with intensity $\lambda$. Then, it can be shown that $\hat{P}(u) \xrightarrow{a.s.} Au =: P(u)$, where $A$ is a constant only depending on $\lambda$.

To construct confidence intervals for $\int_0^t \sigma_s^2 ds$, we need a "studentized" version of the CLT. Note that the limit $\sqrt{2}\int_0^t (P'(u))^{1/2}\sigma_u^2 dW'_u$ in (12) is a mixture normal distribution with mean 0 and variance $2\int_0^t P'(u)\sigma_u^4 du$ (conditional on the data). Similar to the arguments in (8), for any process $A$, a reasonable estimator of $\int_0^t A_u \sigma_u^p du$ is

$$\hat{S}(p, A) := \frac{M^{p/2-1}}{m_p} \sum_{m=1}^{M} \frac{A_{s_{m-2}} |\overline{X}_m - \overline{X}_{m-1}|^p}{w_m^{p/2}}, \qquad \text{for } p \ge 2.$$

THEOREM 2 (Consistent estimation of higher order volatility functional). *For $p \ge 2$ and a continuous function $A_u$, under Assumption 1, as $M \to \infty$, we have*

$$\hat{S}(p, A) \xrightarrow{P} \int_0^t A_u \sigma_u^p du.$$

Using this and the property of the stable convergence, we have the following studentized version of the CLT.

THEOREM 3 (Studentized CLT). *Under Assumption 1, as $M \to \infty$, we have*

$$\frac{\sqrt{M}\left(RV_t^{weight} - \int_0^t \sigma_s^2 ds\right)}{\sqrt{2\hat{S}\left(4, \hat{P}'\right)}} \xrightarrow{\mathcal{S}} \mathcal{N}(0, 1),$$

*where $\hat{P}'(u)$ is a consistent estimator of $P'(u)$.*

**Remark 2.** One obvious choice of $\hat{P}'(u)$ is given by

$$\hat{P}'(u) := [\hat{P}(u) - \hat{P}(u-h)]/h = M\Delta_M^2 h^{-1} \sum_{u-h < s_m \le u} \frac{w_m w_{m+1} + \tilde{v}_{m,1} + \tilde{v}_{m,2}}{w_m w_{m+1}}, \tag{15}$$

for small $h$, which can be shown to converge to $P'(u)$ under appropriate conditions.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

### 3.3. Further discussion for the special case: $L_m \equiv L$

In the special (albeit unrealistic) case of $L_m \equiv L$, many earlier results reduce to much simpler forms. For instance, simple calculation yields $w_m = \frac{2L^2+1}{3L^2}$. It follows from (9)–(10) that

$$RV_t^{weight} = \frac{3L^2}{2L^2+1} RV_t^{ave}. \tag{16}$$

Therefore,

$$\frac{RV_t^{ave}}{RV_t^{weight}} = 1 \qquad \text{if } L = 1,$$
$$< 1 \qquad \text{if } L > 1,$$
$$\to \frac{2}{3} \qquad \text{if } L \to \infty.$$

Since $RV_t^{weight}$ is consistent for $[X, X]_t$ by Theorem 1, the estimator $RV^{ave}$ is always biased downward for $L \geq 2$.

A simple simulation study for the special case of $X = B$ is presented in Table 3 for various values of $L$'s. The results match well with the above analysis and $RV_t^{weight}$ provides excellent estimates of the $IV_t$ for all $L$.

### 3.4. Estimation of the Integrated Quarticity ($IQ_t$)

Another important quantity widely investigated in the financial literature is the integrated quarticity ($IQ_t$):

$$IQ_t := \int_0^t \sigma_s^4 ds, \tag{17}$$

which is useful in constructing a confidence interval for the $IV_t$. Taking $p = 4$ and $A_u \equiv 1$ in Theorem 2, we obtain a consistent estimator of the $IQ_t$ (called the *realized quarticity* ($RQ_t$)):

$$RQ_t^{weight} := \frac{M}{3} \sum_{m=1}^{M} \frac{|\overline{X}_m - \overline{X}_{m-1}|^4}{w_m^2} \xrightarrow{P} \int_0^t \sigma_s^4 ds. \tag{18}$$

**TABLE 3.** The behavior of realized volatility with different $L$

| $L$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $RV_t^{Ave}$ | 1.0001 | .749 | .705 | .686 | .681 | .677 | .673 | .671 | .669 | .674 | .667 |
| $w_m = \frac{2L^2+1}{3L^2}$ | 1 | .75 | .704 | .688 | .68 | .676 | .674 | .672 | .671 | .67 | .668 |
| $RV_t^{weight}$ | 1.0001 | .999 | 1.002 | .999 | 1.002 | 1.002 | .999 | .999 | .997 | 1.006 | .999 |

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

## 4. APPROACH II: RANDOM DRAWING METHOD

### 4.1. Estimation of the $IV_t$

Recall that in $RV_t^{ave}$, we took averages among the multiple transactions $\mathcal{B}_m = \{X_{s_m,1}, ....., X_{s_m,L_m}\}$ for each fixed time $s_m$, resulting in a downward biased estimator. Now, instead of taking averages in the block $\mathcal{B}_m$, we propose to pick a representative from $\mathcal{B}_m$ by drawing randomly with equal probabilities. That is, for each $m = 1, ..., M$, let

$$P(U_m = j) = \frac{1}{L_m}, \qquad j = 1, ..., L_m, \tag{19}$$

i.e., $\{U_m, m = 1, \cdots, M\}$ are i.i.d. r.v.s drawn from the discrete uniform distribution taking values from the set $\{1, 2, \cdots, L_m\}$. Therefore, an estimator of the $IV_t$ is given by

$$RV_t^{RD} := \sum_{m=1}^{M} \left(X_{s_m,U_m} - X_{s_{m-1},U_{m-1}}\right)^2. \tag{20}$$

Is $RV_t^{RD}$ a reasonable estimator of the $IV_t$? Suppose that $X_{s_m,U_m} = X_{R_m}$ for some random $R_m \in (s_{m-1}, s_m]$, which is unobserved. Heuristically, we have (conditional on $\{U_m, m = 1, \cdots, M\}$)

$$X_{s_m,U_m} - X_{s_{m-1},U_{m-1}} = X_{R_m} - X_{R_{m-1}} \approx \sigma_{s_{m-2}} \left(W_{R_m} - W_{R_{m-1}}\right). \tag{21}$$

Then, $RV_t^{RD} \approx \sum_{m=2}^{M} \sigma_{m-2}^2 (R_{m-1} - R_{m-2}) \approx \int_0^t \sigma_s^2 ds$, indicating that $RV_t^{RD}$ should be a reasonable estimator of the $IV_t$. Our next theorem demonstrates that this is indeed the case.

THEOREM 4. *Assumption 1 holds.*

(1). **(Consistency)** *We have*

$$RV_t^{RD} \xrightarrow{P} \int_0^t \sigma_s^2 ds \tag{22}$$

(2). **(CLT)***: We have*

$$\sqrt{M}\left(RV_t^{RD} - \int_0^t \sigma_s^2 ds\right) \xrightarrow{\mathcal{S}} \sqrt{2} \int_0^t \left(Q'_u\right)^{1/2} \sigma_u^2 dW''_u, \tag{23}$$

*where $W''$ is another Wiener process on an extension of $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ independent of $\mathcal{F}$ and $\{U_i, 1 \le i \le M\}$, and we assume $Q(u) = \lim_{M\to\infty} \sum_{s_m \le u} \chi_m \Delta_M$ exists and is differentiable with*

$$\chi_m := \frac{7}{6} + \frac{1}{6L_m^2} + \frac{1}{6L_{m-1}^2} - \frac{1}{2L_m L_{m-1}} + \frac{1}{L_m} - \frac{1}{L_{m-1}}. \tag{24}$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

(3). **(Studentized CLT)**: *We have*

$$\frac{\sqrt{M}\left(RV_t^{RD} - \int_0^t \sigma_s^2 ds\right)}{\sqrt{2\tilde{S}(Q')}} \xrightarrow{\mathcal{S}} \mathcal{N}(0,1), \tag{25}$$

*where,* $\tilde{S}(Q')$ *is defined in (27) below.*

As simple as it is, $RV_t^{RD}$ has some intuitive appeal: random draws introduce some extra variability into our estimator of the $IV_t$, which will correct the bias seen in the $RV_t^{ave}$.

For more stable performance, we can repeat these processes a number of times and take their average as our estimator. Namely,

- for each $b = 1, ..., B$, draw an independent sample $U_1^{(b)}, ....., U_M^{(b)}$ from (19);
- calculate $RV_t^{RD}(b) := \sum_{m=1}^{M} \left(X_{s_m, U_m^{(b)}} - X_{s_{m-1}, U_{m-1}^{(b)}}\right)^2$ and define $\overline{RV}_t^{RD} := B^{-1} \sum_{b=1}^{B} RV_t^{RD}(b)$.

## 4.2. Estimation of the integrated quarticity ($IQ_t$)

We now explore how to estimate $IQ_t := \int_0^t \sigma_s^4 ds$ by the random drawing approach. Heuristically, from (21), we have $(X_{s_m, U_m} - X_{s_{m-1}, U_{m-1}})^4 \approx 3\sigma_{s_{m-2}}^4 (R_m - R_{m-1})^2$. Since $E(R_m - R_{m-1}) = \frac{1}{M} + \frac{1}{2M}(\frac{1}{L_m} - \frac{1}{L_{m-1}})$, and $E(R_m - R_{m-1})^2 = \chi_m \Delta_M^2$ with $\chi_m$ given in (24). This suggests the following estimator:

$$RQ_t^{RD} = \frac{M}{3} \sum_{m=1}^{M} \frac{\left(X_{s_m, U_m} - X_{s_{m-1}, U_{m-1}}\right)^4}{\chi_m}. \tag{26}$$

Using this, an estimator of asymptotic variance of random drawing estimator is:

$$\tilde{S}(Q') = \frac{M}{3} \sum_{m=1}^{M} \frac{Q'(s_{m-2}) \left(X_{s_m, U_m} - X_{s_{m-1}, U_{m-1}}\right)^4}{\chi_m}. \tag{27}$$

## 4.3. Comparison of two estimators

We demonstrate a theoretical comparison of two estimators, $RV_t^{weight}$ and $RV_t^{RD}$, in this section. Because both estimators are consistent and asymptotic normal and achieve the same convergence rate, hence, we compute the asymptotic standard deviations of both. For the purpose of comparison, we investigate the constant case only, $L_m \equiv L$. The asymptotic standard deviation of $RV_t^{weight}$ is given in (12) and (14), that is, $\sqrt{2A \int_0^t \sigma_s^4 ds}$, the asymptotic standard deviation of $RV_t^{RD}$ is given in (23) and (24), it is $\sqrt{2\chi_m \int_0^t \sigma_s^4 ds}$, for constant case,

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

**TABLE 4.** $A$ and $\chi$ with different $L$

| $L$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $A$ | 1 | 1.1389 | 1.1551 | 1.1560 | 1.1539 | 1.1514 | 1.149 | 1.1469 | 1.1451 | 1.1436 | 1.1353 |
| $\chi$ | 1 | 1.125 | 1.1481 | 1.1563 | 1.16 | 1.162 | 1.1633 | 1.1641 | 1.1646 | 1.165 | 1.1663 |

$\chi_m \equiv \chi = \frac{7}{6} - \frac{1}{6L^2}$. Ignoring the common term, we need only to compare $A$ and $\chi$. We display the computation results of $A$ and $\chi$ for different $L$ numerically in Table 4, and graphically in Figure 5.

From Figure 5, we can see that both curves ascend steeply when $L$ changes from 1 to 2, which is because of the multiple transactions effect, then slope gently for larger $L$, and eventually tend to converge. When $L$ increases, $\chi$ continues rising, while $A$ rises in the beginning and falls after having achieved the maximum.

## 5. NUMERICAL EXAMPLES

### 5.1. Data generation

We employ intra-day data throughout the simulations, i.e., we take $t = 1$ day, consisting of 6.5 hours ($= 23,400$ seconds) of trading. We utilize tick-by-tick data, recording observations at every second, so that $M = 23,400$. At each tick, we have $L_i$ ($i = 1, \ldots, M$) multiple transactions, where $L_i$ is independently generated from a Poisson distribution with intensity $\lambda = 2, 4, 5$. Multiple transaction data sets can be generated as follows. We generate an equidistant sample $X_{t_1}, ..., X_{t_n}$

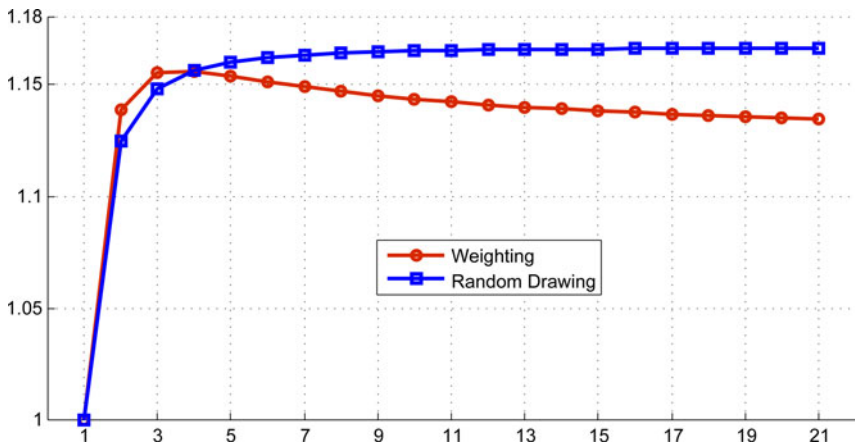



**FIGURE 5.** The asymptotic standard deviations of weighting estimator $RV_t^{weight}$ ($A$) and random drawing estimator $RV_t^{RD}$ ($\chi$) against $L$.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

from the following three commonly employed models, where $n = \sum_{i=1}^{M} L_i$. This sample of size $n$ will then be put into $M = 23,400$ groups sequentially with each group containing $L_i$ multiple transactions, as was performed in (5).

- **Model 1: Geometric Brownian motion (GBM):**

$$dX_t = \mu X_t dt + \sigma X_t dW_t, \qquad \text{with } \mu = 0.05 \text{ and } \sigma = 0.1.$$

- **Model 2. Ornstein-Uhlenbeck (OU) process:**

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t, \qquad \text{with } X_0 = 0, \theta = 2, \mu = 0.05, \sigma = 0.1.$$

- **Model 3. Stochastic volatility Heston model:**

$$dX_t = (\varsigma - K_t/2)dt + \sqrt{K_t}dB_t, \qquad \text{with } X_0 = 0.3, \varsigma = 0.1, \tag{28}$$

where $B$ and $W$ are two Wiener processes with correlation $\rho = -0.7$, and $K_t$ follows a CIR process:

$$dK_t = \alpha(\theta - K_t)dt + \xi\sqrt{K_t}dW_t, \qquad \text{with } K_0 = 0.1, \alpha = 2, \xi = 0.2 \text{ and } \theta = 0.4.$$

### 5.2. Comparisons of the $IV_t$ estimation

For each simulated data, we calculate the three estimators: $RV_t^{weight}$, $RV_t^{RD}$, and $\overline{RV}_t^{RD}$ (the average of 20 repetitions of $RV_t^{RD}$). The experiments are repeated $K = 5,000$ times. We compare these estimators, based on the mean relative bias (i.e., rel.bias $= \frac{1}{K}\sum_{i=1}^{K} \frac{\text{estimator} - IV_t}{IV_t}$), and its standard error. The simulation results including biases and standard errors, are given in Table 5 for three models and three estimators. We display QQ-plots and histograms in Figures 6–8 only for the first estimator due to space limitations because the appearance of them for the other two estimators are similar, which is available upon request. Here is a summary of our simulation results.

- The relative biases in all cases are all very small, ranging from approximately 0.4% to 2%. For all three models, the results given by three estimators are reliable, as they estimate the true values with very small relative biases and standard errors.
- Furthermore, the relative biases for $RV_t^{RD}$ are generally smaller than those for $RV_t^{weight}$. Further, by taking averages, $\overline{RV}_t^{RD}$ can reduce the relative standard error.
- The histograms and Q-Q plots in Figures 6-8 are all in agreement with our theoretical results that the estimators are asymptotically normal distributed.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

**TABLE 5.** Simulation results, where rel.bias $= (RV_t - IV_t)/IV_t$

| | $\lambda$ | $M$ | $RV_t^{weight}$ | $RV_t^{RD}$ | $\overline{RV}_t^{RD}$ |
|---|---|---|---|---|---|
| | | | (rel.bias, rel.s.e.) | (rel.bias, rel.s.e.) | (rel.bias, rel.s.e.) |
| | 2 | 23400 | (0.0216, 0.0107) | (0.0113 0.0105) | (0.0101, 0.0104) |
| Model 1 | 4 | 23400 | (0.0074, 0.0105) | (−0.0052, 0.0108) | (0.0061, 0.0106) |
| | 5 | 23400 | (0.0051, 0.0105) | (0.0053, 0.0107) | (0.0049, 0.0101) |
| | 2 | 23400 | (0.0217, 0.0108) | (0.0201, 0.0104) | (0.0172, 0.0101) |
| Model 2 | 4 | 23400 | (0.0078, 0.0105) | (0.0049, 0.0106) | (0.0058,0.0100) |
| | 5 | 23400 | (0.0048, 0.0102) | (−0.0054, 0.0104) | (−0.0032,0.0102) |
| | 2 | 23400 | (0.0212, 0.0114) | (−0.0206, 0.0114) | (0.0194, 0.0110) |
| Model 3 | 4 | 23400 | (0.0077, 0.0108) | (0.0011, 0.0112) | (0.0015, 0.0111) |
| | 5 | 23400 | (0.0042, 0.0107) | (0.0011, 0.0111) | (0.0006, 0.0109) |

### 5.3. Comparisons of the $IQ_t$ estimation

Next, we consider the estimation of integrated quarticity ($IQ_t$) given in (17): $IQ_t = \int_0^t \sigma_s^4 ds$. In the simulations, we calculate the weighting estimator by

$$RQ_t^{weight} = \frac{M}{3}\sum_{m=1}^{M} \frac{\left|\bar{X}_m - \bar{X}_{m-1}\right|^4}{w_m^2},$$

and random drawing estimator by

$$RQ_t^{RD} = \frac{M}{3}\sum_{m=1}^{M} \frac{\left(X_{s_{m-1},U_m} - X_{s_{m-2},U_{m-1}}\right)^4}{\chi_m}.$$

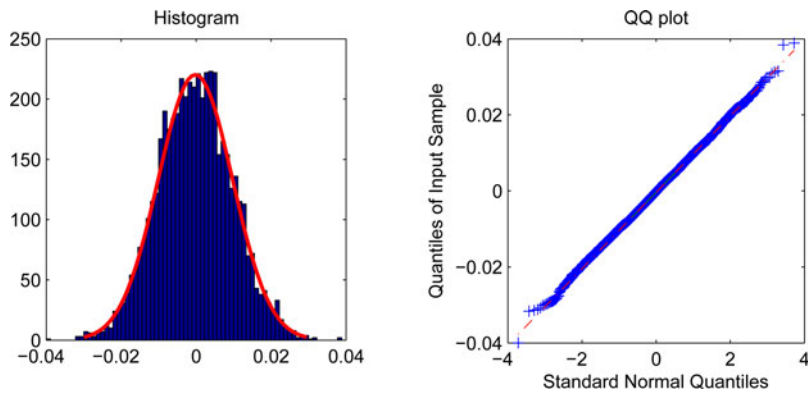



**FIGURE 6.** Histogram and QQ plot of 5,000 values of $\frac{RV_t^{weight} - IV_t}{IV_t}$ in Model 1 for $\lambda = 2$.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

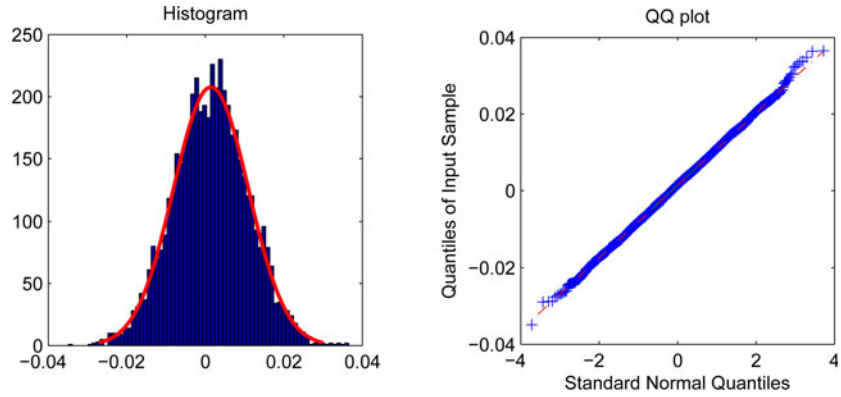



**FIGURE 7.** Histogram and QQ plot of 5,000 values of $\frac{RV_t^{weight} - IV_t}{IV_t}$ in Model 2 for $\lambda = 2$.

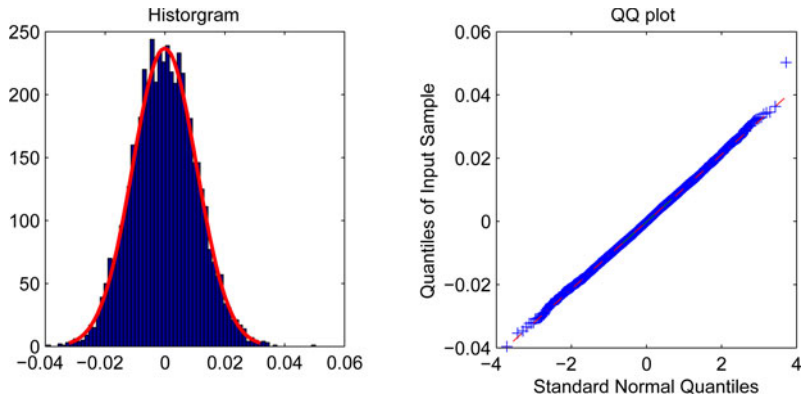



**FIGURE 8.** Histogram and QQ plot of 5,000 values of $\frac{RV_t^{weight} - IV_t}{IV_t}$ in Model 3 for $\lambda = 2$.

Because the simulation results for the three examples are similar, we present only the case for $\lambda = 5$ in Table 6. From Table 6, we see that both the weighting approach and random drawing approach perform very well. Their relative biases for both approaches are very small, all are under 10%.

### 5.4. Estimation under presence of microstructure noise

In the section, we briefly investigate the performance of two types of estimators under the presence of microstructure noise. For simplicity, we generate the data from the following process:

$$X_t = \sigma W_t, \quad Y_{s_i,j} = X_{s_i,j} + \epsilon_{s_i,j}, i = 1, 2, \cdots, M; j = 1, 2, \cdots, L,$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

**TABLE 6.** Relative biases and standard errors of $RQ_t^{weight}$ and $RQ_t^{RD}$

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| | (rel.bias, rel.s.e.) | (rel.bias, rel.s.e.) | (rel.bias, rel.s.e.) |
| $RQ_t^{weight}$ | (0.0711, 0.0213) | (0.0712, 0.0291) | (0.0891, 0.0312) |
| $RQ_t^{RD}$ | (0.0793, 0.0237) | (0.0701, 0.0298) | (0.0913, 0.0372) |

where, $W_t$ is a standard Brownian motion. The parameter values in the model are chosen to be close to those utilized in Jacod, Li, Mykland, Podolskij, and Vetter (2009); i.e., $\sigma = 0.2/\sqrt{252}$, $\epsilon \sim^{iid} N(0, \omega^2)$, where $\omega$ will be specified later, and $M = 23,400$. The number of transactions within a second is $L$. That is, we first generate $M \times L$ discretized points of $X_t$, then group them according to the observation scheme in Figure 1 of page 3. Based on those averages $\{\bar{Y}_i, i = 1, 2, \ldots, M\}$, we employ the pre-averaging approach to construct an estimator for the integrated volatility as follows:

$$\widehat{IV}_t = \frac{1}{\bar{g}(2)}\left(\frac{1}{k_M}\sum_{i=0}^{M-k_M}\left(\Delta_{i,k_M}Y\right)^2 - \frac{\bar{g}'(2)}{2M}\sum_{i=1}^{M}\left(\bar{Y}_i - \bar{Y}_{i-1}\right)^2\right),$$

where, $K_M = \lfloor\sqrt{M}\rfloor$, and we choose $g(x) = \min(x, 1-x)$, hence $\bar{g}(2) = \frac{1}{12}$ and $\bar{g}'(2) = 1$. The pre-averaging estimator for the random drawing scheme is the same as the original pre-averaging estimator because we only have one piece of data at each time point. We compare the performance of two pre-averaging estimators (averaging+weighting+pre-averaging and random-drawing+pre-averaging, AWP estimator and RDP estimator afterward, respectively). In the simulation, we let the magnitude of microstructure noise vary, $\omega$ changes from 0.0005 to 0.005 with a uniform step $\delta = 0.0001$, that is, we have a total of 46 averaged relative biases and standard deviations for each estimator by repeating the simulation 1,000 times. We display the results of $L = 2$ and $L = 5$ in Figure 9 and Figure 10, respectively.

From Figures 9–10, we can see that both estimators perform well because finite sample biases and finite sample variances are small, nevertheless, they grow when the variance of noise increases. The finite sample bias and finite sample variance of the AWP estimator for all cases is smaller than that of the RDP estimator. This becomes obvious when $\omega$ is large. For both cases when $L = 2$ and $L = 5$, the RDP estimator is very sensitive to the magnitude of the microstructure noise, whereas the AWP estimator is less sensitive than the RDP estimator; it is particularly robust when $L = 5$. This is reasonable because the AWP estimator smoothes the data two times (averaging and pre-averaging), hence, the effect of microstructure noise is significantly reduced, compared with the RDP estimator. The results demonstrate that the AWP estimator outperforms the RDP estimator, so we recommend employing the AWP estimator when microstructure noise is present, in

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

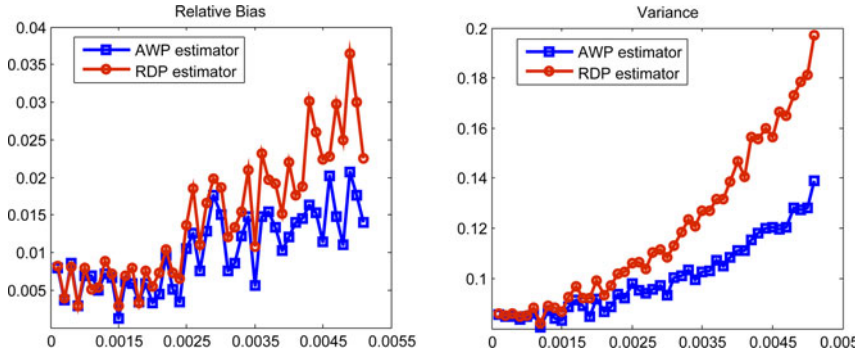



**FIGURE 9.** The $x$-axis is $\omega$ (the standard deviation of microstructure noise). $L = 2$.

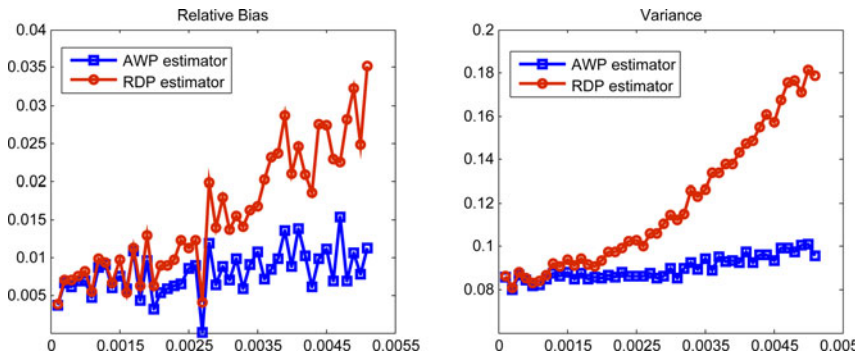



**FIGURE 10.** The $x$-axis is $\omega$ (the standard deviation of microstructure noise). $L = 5$.

particular, when $L$ is large. We leave the detailed proof of the theoretical aspects to future work.

## 6. APPLICATION TO FINANCIAL TIME SERIES

In this section, we apply the estimation procedures to some real data. They consist of intraday data of the three most actively traded stocks: Microsoft (MSFT), Intel (INTC) and Pfizer (PFE) on August 1, 2000. The data were obtained from the TAQ2 data base. Table 7 presents the total numbers of transactions and of observed non-zero intervals. Clearly, Intel is the most actively traded stock, followed by Microsoft, then Pfizer.

The $IV_t$ of the log price is estimated by the proposed methods given in the paper, i.e, $RV_t^{weight}$ and $\overline{RV}_t^{RD}$ (the average of 20 $RV_t^{RD}s$). In addition, the $RV_t^{ave}$ is included for comparisons. The $RV_t^{RD}$ changes much for different samples, however, $\overline{RV}_t^{RD}$ is quite stable when $B > 10$. Hence, we do not include the

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

**TABLE 7.** The real data

| | Pfizer | Microsoft | Intel |
|---|---|---|---|
| Total number of transactions | 6434 | 18501 | 39745 |
| Observed non-zero intervals | 5096 | 10174 | 15506 |
| Average number of trans. per interval | 1.3 | 1.8 | 2.6 |

$RV_t^{RD}$ curve in the picture. The results are presented in Figure 11. Here are some observations from the results.

- For the three cases, we have $RV_t^{ave} \le RV_t^{weight} \le \overline{RV}_t^{RD}$. The first inequality is to be expected since our earlier simulation studies demonstrate that $\overline{RV}_t^{ave}$ underestimates the $IV_t$ while $RV_t^{weight}$ and $\overline{RV}_t^{RD}$ are consistent estimators.
- The $RV_t^{weight}$ curve and the $\overline{RV}_t^{RD}$ curve are quite close to each other for the Pfizer and Microsoft data sets, relative to $RV_t^{ave}$. However, for the Intel data set, they exhibit some appreciable differences.
- The differences between $RV_t^{weight}$ and $\overline{RV}_t^{RD}$ may be due to the microstructure noise in the real data. First, the higher the frequency is, the more serious the effect of microstructure noise is. Second, $RV_t^{weight}$ uses local smoothing to reduce the effect of microstructure noise; the more transactions per non-zero interval we have, the more reductions we will obtain. Finally, $\overline{RV}_t^{RD}$ keeps the microstructure noise more or less untreated. These help to explain why the difference between $RV_t^{weight}$ and $\overline{RV}_t^{RD}$ is the largest for Intel, which has the highest frequency data and also the largest number of transactions per interval (see Table 7).

## 7. SUMMARY AND SOME OPEN PROBLEMS

Multiple transactions occur commonly for high-frequency data, particularly for heavily traded stocks. In the context of estimation of integrated volatility, we pro-

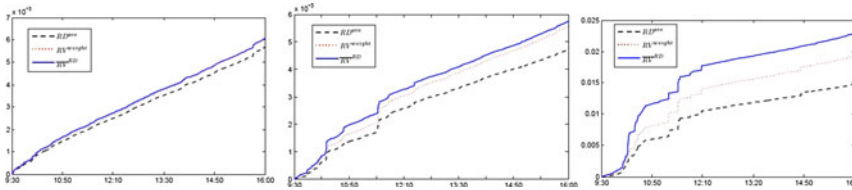


**FIGURE 11.** The daily $RV_t^{ave}$, $RV_t^{weight}$ and $\overline{RV}_t^{RD}$ curves. Left: Pfizer; Middle: Microsoft; Right: Intel.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

pose two easy-to-use approaches to address this phenomenon: weighting-averages and random drawing. They have been shown to be effective both in theory and in simulation studies. The weighting-averages estimator $RV_t^{weight}$ employs local averages of increments, thus having a similarity with Jacod et al. (2009), although they employ a local average to address the microstructure noise effect. This is also shown in our proofs. We are combining both methods (pre-averaging and weighting) to address the microstructure noise and multiple transactions simultaneously, but the work is ongoing.

The present work only represents the first step in this direction. There are many issues involving high-frequency data. Here are some of them.

- **Microstructure noise.** It is widely accepted that the observed prices are contaminated by microstructure noise, due to bid-ask spreads, rounding errors, etc. The $IV_t$ estimation in the presence of microstructure noise has also been considered by various authors e.g., Zhang, Mykland, and Aït-Sahalia (2005), Zhang (2006), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a; 2008b), Barndorff-Nielsen et al. (2008b), Podolskij and Vetter (2009), Jacod, Podolskij and Vetter (2009) and Xiu (2010), among others. We briefly discuss the estimation of integrated volatility under simultaneous presence of multiple observations and microstructure noise in the current work by a simulation study, the detailed theoretical aspects of which are to be investigated.

- **Price jumps.** Empirical and statistical evidence of asset price discontinuities or "jumps" in financial markets has been well documented in the literature. Estimation of the $IV_t$ in the presence of the jumps has also been considered by various authors e.g., Barndorff-Nielsen and Shephard (2006), Mancini (2009), and Jacod, Podolskij and Vetter (2010) among others.

- **Non-equidistance, random sampling, endogeneity, etc.** Much of the literature on high-frequency financial data assumes that the observations are equidistant and deterministic. However, in reality, transaction times might be non-equidistant and random and include the presence of endogeneity in the times. These issues have also been addressed by Aït-Sahalia and Mykland (2003), Aït-Sahalia and Mykland (2004), Mykland and Zhang (2006), Barndorff-Nielsen et al. (2008a), Hayashi, Jacod, and Yoshida (2011), and Li, Mykland, Renault, Zhang, and Zheng (2014), among others.

Therefore, the $IV_t$ estimation can be conducted under one or more of the above frameworks, as well as the presence of multiple transactions. Some of these are currently under our investigation.

So far, we have focused on estimating the integrated volatility $IV_t$ with multiple transactions for high-frequency data in this paper. However, we might also be interested in studying other characteristics of the asset price processes in the presence of multiple transactions. For instance, we can investigate the estimation

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

of the *jump activity index* (as in Mykland and Zhang (2009), Zhang (2011), Aït-Sahalia et al. (2011), Mykland and Zhang (2012), Aït-Sahalia and Jacod (2009), Jing et al. (2012), Todorov and Bollerslev (2010), and in Todorov and Tauchen (2010)) or covariance matrix (as in Barndorff-Nielsen and Shephard (2004)) when multiple transactions occur.

## 8. THE TECHNICAL PROOFS

We recall and denote some notations needed in the proofs:

$$\Delta_M = s_m - s_{m-1},\ \ \Delta_m^n = \Delta_M / L_m,\ \ \Delta_i^n t = t_i - t_{i-1},\ \ \Delta_i^n X = X_{t_i} - X_{t_{i-1}},\ \ \mathcal{G}_m = \mathcal{F}_{s_{m-1}}. \tag{29}$$

When $L_m \equiv L$, we simply write $\Delta_n = \Delta_i^n t$. And for a process $\{Z(t), t \geq 0\}$, we denote

$$V_t^M(Z,p) := \frac{M^{p/2-1}}{m_p} \sum_{m=1}^{M} \frac{\left|\overline{Z}_m - \overline{Z}_{m-1}\right|^p}{w_m^{p/2}}. \tag{30}$$

When $Z = X$ and $p = 2$, it is just $RV_t^{weight}$. Throughout the proof, $C$ denotes a constant that may change from line to line. By using a localization procedure as described in Jacod (2012), we assume that all the coefficients related to $X$ are bounded in the interval $[0,t]$. Let $X = Dr + Di$, where $Dr(t) = \int_0^t b_s ds$ and $Di(t) = \int_0^t \sigma_s dW_s$. First, we show that the drift term $Dr(t) = \int_0^t b_s ds$ can be ignored when we consider the asymptotic behavior (consistency and asymptotic normality) of $V_t^M(X,p)$, $p > 0$. We show this by considering the following two cases:

1. $p \leq 1$. By the elementary inequality $||x+y|^p - |x|^p| \leq |y|^p$, we have

$$\begin{aligned}\left|V_t^M(Dr+Di,p) - V_t^M(Di,p)\right| &\leq CM^{p/2-1} \sum_{m=1}^{M} \left| \sum_{k=1}^{L_{m-1}+L_m} g_{m,k}\, \Delta_{N_{m-2}+k}^n Dr \right|^p \\ &\leq CM^{p/2} \sup_{1\leq m\leq M} \Big| \sum_{k=1}^{L_{m-1}+L_m} g_{m,k}\, \Delta_{N_{m-2}+k}^n Dr \Big|^p.\end{aligned}$$

   Using the boundedness and continuity of $\{b_s, s \geq 0\}$ and boundedness of $g_{m,k}$, one can easily obtain

$$\limsup_{M\to\infty} |V_t^M(Dr+Di,p) - V_t^M(Di,p)| \to 0. \tag{31}$$

2. $p > 1$. We have the inequality $\left||x+y|^p - |x|^p\right| \leq C_p\left(|y|^p + |x|^{p-1}|y|\right)$.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

We almost surely have

$$\begin{aligned}
&\left|V_t^M(Dr+Di,p)-V_t^M(Di,p)\right| \\
&\quad\le C_p\Delta_M^{p/2}+C_p\sup_{1\le m\le M}\left(M^{p/2}\left|\sum_{k=1}^{L_{m-1}+L_m}g_{m,k}\Delta^n_{N_{m-2}+k}Dr\right|\right. \\
&\qquad\qquad\left.\times\left|\sum_{k=1}^{L_{m-1}+L_m}g_{m,k}\Delta^n_{N_{m-2}+k}Di\right|^{p-1}\right) \\
&\quad\le C_p\Delta_M^{p/2}+C_p\sup_{1\le i\le n}\left(M^{p/2-1}\left|\Delta_i^nDi\right|^{p-1}\right) \\
&\quad\le C_p\Delta_M^{p/2}+C_pM^{p/2-1}\sup_{1\le i\le n}\left(\left(\Delta_i^n\cdot\log\left(\frac{1}{\Delta_i^n}\right)\right)^{(p-1)/2}\right) \\
&\quad\le C_p\Delta_M^{p/2}+C_p\sup_{1\le i\le n}\left(\sqrt{\Delta_i^n}\cdot\left(\log\left(\frac{1}{\Delta_i^n}\right)\right)^{(p-1)/2}\right),
\end{aligned}$$

where, we use the modulus of continuity of Brownian motion (see Theorem 1.12 of Mörters and Peres (2010)) and boundedness of $L_m$. Hence by L'Hôpital's rule (after taking logarithms), we have

$$\limsup_{M\to\infty}\left|V_t^M(Dr+Di,p)-V_t^M(Di,p)\right|\to 0. \tag{32}$$

For the central limit theorem, we restrict attention to the case of $p=2$ only, which can be extended to more general cases, see Barndorff-Nielsen et al. (2006), for example. Observe that

$$\sqrt{M}\left|V_t^M(Dr+Di,2)-V_t^M(Di,2)\right|=I_M+II_M,$$

where

$$\begin{aligned}
I_M&=\sqrt{M}\sum_{m=1}^{M}\left(\sum_{k=1}^{L_{m-1}+L_m}g_{m,k}\Delta^n_{N_{m-2}+k}Dr\right)^2\Big/\omega_m, \\
II_M&=2\sqrt{M}\sum_{m=1}^{M}\left(\sum_{k=1}^{L_{m-1}+L_m}g_{m,k}\Delta^n_{N_{m-2}+k}Dr\sum_{k=1}^{L_{m-1}+L_m}g_{m,k}\Delta^n_{N_{m-2}+k}Di\right)\Big/\omega_m.
\end{aligned}$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

One can easily show that $\lim_{M\to\infty} I_M \xrightarrow{P} 0$. Now we denote

$$III_M = 2\sqrt{M}\sum_{m=1}^{M}\left(b_{s_{m-2}}\sigma_{s_{m-2}}\sum_{k=1}^{L_{m-1}+L_m} g_{m,k}\Delta^n_{N_{m-2}+k} \times \sum_{k=1}^{L_{m-1}+L_m} g_{m,k}\Delta^n_{N_{m-2}+k}W\right)\Big/\omega_m =: \sum_{m=1}^{M}\zeta_m.$$

On the one hand, we can show that $\lim_{M\to\infty}(II_M - III_M) \xrightarrow{P} 0$ by using Hölder's inequality (one may need to plug in one term). On the other hand, observing that $\zeta_m$ is $\mathcal{G}_{m+1}$-measurable, in view of Lemma 5.2 of Barndorff-Nielsen et al. (2006), it suffices to show that

$$\sum_{m=1}^{M} E\left[(\zeta_m)^2\Big|\mathcal{G}_{m-1}\right] \xrightarrow{P} 0,$$

because $\sum_{m=1}^{M} E\left[\zeta_m\Big|\mathcal{G}_{m-1}\right] = 0$. Now, when $n \to \infty$,

$$\sum_{m=1}^{M} E\left[(\zeta_m)^2\Big|\mathcal{G}_{m-1}\right] = 4M\sum_{m=1}^{M}\left(b^2_{s_{m-2}}\sigma^2_{s_{m-2}}\left(\sum_{k=1}^{L_{m-1}+L_m} g_{m,k}\Delta^n_{N_{m-2}+k}\right)^2\Delta_M\right)\Big/\omega_m \xrightarrow{P} 0.$$

Then, by Lemma 5.2 of Barndorff-Nielsen et al. (2006), we have $\lim_{M\to\infty}\sum_{m=1}^{M}\zeta_m \xrightarrow{P} 0$. Hence, in the proof of the main theorems, without loss of generality, we set the drift term of $X$ to be zero.

**Proof of Theorem 1.**

(1) **(Consistency)**

First, we prove the theorem for the case of $L_m \equiv L$ for a given positive integer $L$. In this case, we have

$$w_m \equiv w = \frac{1}{L_{m-1}}\sum_{k=1}^{L_{m-1}} g^2_{2,k} + \frac{1}{L_m}\sum_{k=L_{m-1}+1}^{L_{m-1}+L_m} g^2_{2,k} = \frac{2L^2+1}{3L^2},$$

for all $1 \le m \le M$. Simple computation shows that

$$\sum_{m=1}^{M}\left(\bar{X}_{s_m} - \bar{X}_{s_{m-1}}\right)^2 = \sum_{m=1}^{M}\left(\frac{1}{L}\sum_{j=1}^{L} X_{t_{(m-1)L+j}} - \frac{1}{L}\sum_{j=1}^{L} X_{t_{(m-2)L+j}}\right)^2$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

$$= \sum_{m=1}^{M} \left( \sum_{k=1}^{2L} g_2(\frac{k-1}{2L}) \Delta^n_{(m-2)L+k} X \right)^2$$
$$=: \sum_{m=1}^{M} \left( \tilde{X}_m \right)^2, \tag{33}$$

where for convenience, we take $X_{t_i} = 0$ when $i \leq 0$. Observe that $\tilde{X}_m = \int_{s_{m-2}}^{s_m} g_2\left(\frac{\lfloor (s-s_{m-2})/(2\Delta_M) \rfloor}{2L}\right) dX_s$, $\lfloor A \rfloor$ is the integer part of a real number $A$. By Itô's formula, one can obtain $(\tilde{X}_m)^2 = \eta_m + \eta'_m$ with

$$\eta_m = \int_{s_{m-2}}^{s_m} g_2^2 \left( \frac{\lfloor (s-s_{m-2})/(2\Delta_M) \rfloor}{2L} \right) \sigma_s^2 ds,$$
$$\eta'_m = 2 \int_{s_{m-2}}^{s_m} X\left(n,m\right)_s dX\left(n,m\right)_s, \tag{34}$$

where

$$X(n,m)_s = \int_{s_{m-2}}^{s} g_2 \left( \frac{\lfloor (u-s_{m-2})/(2\Delta_M) \rfloor}{2L} \right) dX_u, \tag{35}$$

for $s_{m-2} \leq s \leq s_m$. Rewriting (33) with (34), we obtain

$$\sum_{m=1}^{M} \eta_m = w \int_0^t \sigma_s^2 ds + C_M + D_M - w \int_{s_M}^t \sigma_s^2 ds, \tag{36}$$

where

$$\begin{cases} C_M = \sum_{m=1}^{M} \int_{s_{m-2}}^{s_m} g_2^2 \left( \frac{\lfloor (s-s_{m-2})/(2\Delta_M) \rfloor}{2L} \right) \left( \sigma_s^2 - \sigma_{s_{m-2}}^2 \right) ds, \\ D_M = -\sum_{m=1}^{M} w \int_{s_{m-2}}^{s_m} (\sigma_s^2 - \sigma_{s_{m-2}}^2) ds. \end{cases}$$

We prove $C_M \xrightarrow{P} 0$ next. Employing Cauchy–Schwarz inequality, we have

$$E|C_M| \leq \sum_{m=1}^{M} E \left| \int_{s_{m-2}}^{s_m} g_2^2 \left( \frac{\lfloor (s-s_{m-2})/(2\Delta_M) \rfloor}{2L} \right) \left( \sigma_s^2 - \sigma_{s_{m-2}}^2 \right) ds \right| \tag{37}$$
$$\leq \sum_{m=1}^{M} \int_{s_{m-2}}^{s_m} g_2^2 \left( \frac{\lfloor (s-s_{m-2})/(2\Delta_M) \rfloor}{2L} \right) \sqrt{E\left| \sigma_s^2 - \sigma_{s_{m-2}}^2 \right|^2} ds.$$

Consequently,

$$E|C_M| \leq C\sqrt{\Delta_n} \to 0,$$

when $\Delta_n \to 0$. Therefore, Markov's inequality yields $C_M \xrightarrow{P} 0$. A similar result holds for $D_M$. Hence, we have

$$\sum_{m=1}^{M} \eta_m \xrightarrow{P} w \int_0^t \sigma_s^2 ds. \tag{38}$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

To prove the theorem, we then need to prove that

$$\sum_{m=1}^{M} \eta'_m \xrightarrow{P} 0. \tag{39}$$

Note that, for $j \geq 2$ and $1 \leq m \leq M$,

$$E\left[\eta'_m \eta'_{m+j}\right] = 0, \quad E\left[(\eta'_m)^2\right] \leq 4E\left[\int_{s_{m-2}}^{s_m} \sigma_s^2 ds \cdot \sup_{s\in[s_{m-2},s_m]} X^2(n,m)_s\right]. \tag{40}$$

We utilize (40) to obtain

$$\begin{aligned} E\left[\sum_{m=1}^{M} \eta'_m\right]^2 &\leq 3\sum_{m=1}^{M} E\left[\eta'_m\right]^2 \\ &\leq CE\left[\int_0^t \sigma_s^2 ds \cdot \sup_{1\leq m\leq M} \sup_{s\in[s_{m-2},s_m]} X^2(n,m)_s\right] \to 0. \end{aligned}$$

Hence, combining this with (38) together yields

$$RV_t^{weight} = \sum_{m=1}^{M} \frac{\left(\bar{X}_m - \bar{X}_{m-1}\right)^2}{w} \xrightarrow{P} \int_0^t \sigma_s^2 ds. \tag{41}$$

Therefore, we have finished the proof of consistency for the case of $L_m \equiv L$.

For the general case, denote

$$g_m(s) = \begin{cases} \frac{\lfloor (s-s_{m-2})/\Delta_{m-1}^n \rfloor}{L_{m-1}}, & \text{if } s \in (s_{m-2}, s_{m-1}) \\ 1 - \frac{\lfloor (s-s_{m-1})/\Delta_m^n \rfloor}{L_m}, & \text{if } s \in [s_{m-1}, s_m]. \end{cases} \quad \text{and}$$

$$\sum_{m=1}^{M} \frac{\left(\bar{X}_m - \bar{X}_{m-1}\right)^2}{w_m} =: \sum_{m=1}^{M} \left(\eta''_m + \eta'''_m\right),$$

with

$$\eta''_m = \frac{1}{w_m} \int_{s_{m-2}}^{s_m} g_m^2(s)\sigma_s^2 ds, \quad \eta'''_m = \frac{2}{w_m} \int_{s_{m-2}}^{s_m} X'(n,m)_s dX'(n,m)_s,$$

where, $X'(n,m)_s = \int_{s_{m-2}}^{s} g_m(u) dX_u$. By noting that $\int_{s_{m-2}}^{s_m} g_m^2(s) ds = \omega_m \Delta_M$, we obtain a new decomposition

$$\sum_{m=1}^{M} \eta''_m = \int_0^t \sigma_s^2 ds + C'_M + D'_M - \int_{s_M}^t \sigma_s^2 ds, \tag{42}$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

where

$$C'_M = \sum_{m=1}^{M} \frac{1}{w_m} \int_{s_{m-2}}^{s_m} g_m^2(s)(\sigma_s^2 - \sigma_{s_{m-2}}^2)ds, \qquad D'_M = -\sum_{m=1}^{M} \int_{s_{m-2}}^{s_m} (\sigma_s^2 - \sigma_{s_{m-2}}^2)ds.$$

$C'_M \xrightarrow{P} 0$ and $D'_M \xrightarrow{P} 0$ can be proved similarly.

(2) **(Central limit theorem)**.

We first show the proof of special case $L_m \equiv L$. The proof in the general case will be provided in Remark 3 which is yet to come. By employing (33)–(36), to prove the theorem, it is enough to show that

$$\sqrt{M} C_M/\omega \xrightarrow{P} 0, \quad \sqrt{M} D_M/\omega \xrightarrow{P} 0,$$
$$\sqrt{M}\left(\sum_{m=1}^{M} \frac{\eta'_m}{\omega} - \int_0^t \sigma_s^2 ds\right) \xrightarrow{\mathcal{S}} \sqrt{2A} \int_0^t \sigma_s^2 dW'_s, \tag{43}$$

where $A$ is defined in Remark 1. We show $\sqrt{M} D_M \xrightarrow{P} 0$ first. Let

$$\begin{aligned}\gamma_m &:= \sqrt{M} w \int_{s_{m-2}}^{s_m} \left(\sigma_s^2 - \sigma_{s_{m-2}}^2\right) ds \\ &= \sqrt{M} w \int_{s_{m-2}}^{s_m} 2\sigma_{s_{m-2}}(\sigma_s - \sigma_{s_{m-2}})ds + \sqrt{M} w \int_{s_{m-2}}^{s_m} \left(\sigma_s - \sigma_{s_{m-2}}\right)^2 ds \\ &=: \tilde{\gamma}_m + \tilde{\chi}_m.\end{aligned}$$

We decompose $\tilde{\gamma}_m$ further as $\tilde{\gamma}_m = \gamma'_m + \gamma''_m + \gamma'''_m$ with

$$\gamma'_m = \sqrt{M} \int_{s_{m-2}}^{s_m} 2\sigma_{s_{m-2}} \int_{s_{m-2}}^{s} b''_u du ds,$$
$$\gamma''_m = \sqrt{M} \int_{s_{m-2}}^{s_m} 2\sigma_{s_{m-2}} \left(\int_{s_{m-2}}^{s} \sigma'_u dW_u + \int_{s_{m-2}}^{s} V_u dB_u\right) ds,$$
$$\gamma'''_m = \sqrt{M} \int_{s_{m-2}}^{s_m} 2\sigma_{s_{m-2}} \left(\int_{s_{m-2}}^{s} \int_E \phi \circ \kappa\left(u-, x\right)\left(\underline{\mu} - \underline{\nu}\right)\left(du, dx\right)\right) ds,$$

where $b'' = b' + \int_E \left(\kappa - \phi \circ \kappa\left(u-, x\right)\right)\underline{\nu}\left(dx\right)$. Employing Cauchy-Schwarz inequality and boundedness of $b'$ yields

$$E|\gamma'_m| \leq \frac{C}{M^{3/2}}. \tag{44}$$

On the other hand, we have $E[\gamma''_m|\mathcal{G}_{m-1}] = 0$. Apply Hölder's and Burkholder's inequalities and boundedness to obtain

$$E\left[(\gamma''_m)^2|\mathcal{G}_{m-1}\right] \leq \frac{C}{M}\left(E\left(M \int_{s_{m-2}}^{s_m} \int_{s_{m-2}}^{s} \sigma'_u dW_u ds\right)^2\right.$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

$$
\begin{aligned}
&+E\left(M\int_{s_{m-2}}^{s_m}\int_{s_{m-2}}^{s} V_u dB_u ds\right)^2\Bigg)\\
&\le CE\left[\int_{s_{m-2}}^{s_m}\left(\int_{s_{m-2}}^{s}\sigma_u' dW_u\right)^2 ds+\int_{s_{m-2}}^{s_m}\left(\int_{s_{m-2}}^{s} V_u dB_u\right)^2 ds\right]\\
&\le \frac{C}{M^2}. \qquad \textbf{(45)}
\end{aligned}
$$

Similarly, we have

$$
E\left[(\gamma_m''')^2\Big|\mathcal{G}_{m-1}\right]\le\frac{C}{M^2}. \qquad \textbf{(46)}
$$

With Itô's Isometry, we can show $E|\tilde{\chi}_m|\le\frac{C}{M^{3/2}}$. Combining (44)–(46) and noting that

$$
\sqrt{M}D_M=-\sum_{m=1}^{M}(\tilde{\gamma}_m+\tilde{\chi_m}),
$$

and $E\left[\gamma_m''\gamma_{m+j}''\right]=0$, $E\left[\gamma_m'''\gamma_{m+j}'''\right]=0$ for $j\ge 2$ and $1\le m\le M$, we have shown that $\sqrt{M}D_M\xrightarrow{P}0$. Similarly, we can prove that $\sqrt{M}C_M\xrightarrow{P}0$.

Next we show that $\sqrt{M}\left(\sum_{m=1}^{M}\frac{\eta_m'}{\omega}-\int_0^t\sigma_s^2 ds\right)\xrightarrow{\mathcal{S}}\sqrt{2A}\int_0^t\sigma_s^2 dW_s'$. Letting $\eta_m'':=\sqrt{M}\left(\frac{\eta_m'}{\omega}-\int_{s_{m-1}}^{s_m}\sigma_s^2 ds\right)$, we simply have

$$
E\left[\eta_m''\Big|\mathcal{G}_{m-1}\right]=0,\quad E\left[(\eta_m'')^4\Big|\mathcal{G}_{m-1}\right]\le C/M^4. \qquad \textbf{(47)}
$$

Then, we find the variance. Instead of calculating from $\eta_m'$ directly, we compute it from $\left(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\right)^2$. Indeed, they are asymptotically the same one in view of (34). By (8), we have

$$
var\left[\left(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\right)^2\Big|\mathcal{G}_{m-1}\right]=2\sigma_{s_{m-2}}^4\omega^2\Delta_M^2+o_P\left(\Delta_M\right). \qquad \textbf{(48)}
$$

Let $w=\frac{1}{L}\sum_{k=1}^{L-1}\left(\frac{k}{L}\right)^2+\frac{1}{L}\sum_{j=1}^{L}\left(1-\frac{j-1}{L}\right)^2=:w_1+w_2$. Some easy computations show that

$$
\begin{aligned}
&cov\left[\left(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\right)^2,\left(\bar{X}_{s_{m+1}}-\bar{X}_{s_m}\right)^2\Big|\mathcal{G}_{m-1}\right]\\
&=E\left[\left(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\right)^2(\bar{X}_{s_{m+1}}-\bar{X}_{s_m})^2\Big|\mathcal{G}_{m-1}\right]-\sigma_{s_{m-2}}^4\omega^2\Delta_M^2+o_P\left(\Delta_M^2\right)\\
&=\sigma_{s_{m-2}}^4\left(w_1^2+w_1w_2+w_2^2+\tilde{w}^2-\omega^2\right)\Delta_M^2+o_P\left(\Delta_M^2\right)\\
&=\sigma_{s_{m-2}}^4\left(\tilde{w}^2-w_1w_2\right)\Delta_M^2+o_P\left(\Delta_M^2\right),
\end{aligned}
$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

where, $\tilde{w}^2 = \frac{1}{L^6}\left\{\sum_{k=1}^{L-1}\sum_{j=0}^{k}\left(k^2(L-j)^2+(L-k)^2j^2+4kj(L-k)(L-j)\right)\right\}$. We get

$$\sum_{m=1}^{M} E\left[(\eta_m'')^2\middle|\mathcal{G}_{m-1}\right]+2\sum_{m=1}^{M-1} E\left[\eta_m''\eta_{m+1}''\middle|\mathcal{G}_{m-1}\right]$$
$$=\sum_{m=1}^{M} 2\sigma_{s_{m-2}}^4\left(1+\frac{\tilde{w}^2}{w^2}-\frac{w_1w_2}{w^2}\right)\Delta_M+o_P\left(\Delta_M\right)\to 2A\int_0^t\sigma_s^4ds. \tag{49}$$

See (14) for the description of $A$. For any bounded martingale $N$ on $(\Omega,\mathcal{F},\mathcal{G}_t,P)$, which is orthogonal to $W$,

$$E\left[\eta_m''\Delta_m^M N|\mathcal{G}_{m-1}\right]=0, \tag{50}$$

where $\Delta_m^M = N_{s_m}-N_{s_{m-2}}$, it can be shown in the same way as in Proposition 4.1 of Barndorff-Nielsen et al. (2006). Combining (47)–(50), we deduce that

$$\sqrt{M}\left(\sum_{m=1}^{M}\frac{\eta_m''}{w}-\int_0^t\sigma_s^2ds\right)\xrightarrow{\mathcal{S}}\sqrt{2A}\int_0^t\sigma_s^2dW_s' \tag{51}$$

from Theorem IX. 7.28 of Jacod and Shiryayev (2003) and Theorem 2.4 of Barndorff-Nielsen et al. (2006). This completes the proof. ■

**Remark 3.** For the general case, where $L_m$ depends on $m$, we only need to repeat the previous proof, except for calculating the variance, which is given in (13). We present the details here. To compute the variance, let

$$w_m=\frac{1}{L_{m-1}}\sum_{k=1}^{L_{m-1}-1}\left(\frac{k}{L_{m-1}}\right)^2+\frac{1}{L_m}\sum_{j=1}^{L_m}\left(1-\frac{j-1}{L_m}\right)^2=:w_{m,1}+w_{m,2}.$$

First, we have

$$var\left[\left(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\right)^2\middle|\mathcal{G}_{m-1}\right]=2\sigma_{s_{m-2}}^4w_m^2\Delta_M^2+o_P\left(\Delta_M^2\right), \tag{52}$$

where $o_P\left(\Delta_M^2\right)$ is from the approximation (8). Second,

$$cov\left[\left(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\right)^2,\left(\bar{X}_{s_{m+1}}-\bar{X}_{s_m}\right)^2\middle|\mathcal{G}_{m-1}\right]$$
$$=E\left[\left(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\right)^2\left(\bar{X}_{s_{m+1}}-\bar{X}_{s_m}\right)^2\middle|\mathcal{G}_{m-1}\right]$$
$$-E\left[\left(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\right)^2\middle|\mathcal{G}_{m-1}\right]E\left[\left(\bar{X}_{s_{m+1}}-\bar{X}_{s_m}\right)^2\middle|\mathcal{G}_{m-1}\right].$$

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

It is easy to show that

$$E\Big[\Big(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\Big)^2\Big|\mathcal{G}_{m-1}\Big]=\sigma^2_{s_{m-2}}w_m+o_P\Big(\Delta_M^2\Big),\quad E\Big[\Big(\bar{X}_{s_{m+1}}-\bar{X}_{s_m}\Big)^2\Big|\mathcal{G}_{m-1}\Big]$$
$$=\sigma^2_{s_{m-2}}w_{m+1}+o_P\Big(\Delta_M^2\Big). \tag{53}$$

Let

$$\begin{cases}\alpha_m=\sigma_{s_{m-2}}\sum_{k=1}^{L_{m-1}}\frac{k-1}{L_{m-1}}\Delta^n_{N_{m-2}+k}W,\ \beta_m=\sigma_{s_{m-2}}\sum_{k=1}^{L_m}\Big(1-\frac{k-1}{L_m}\Big)\Delta^n_{N_{m-1}+k}W,\\ \alpha'_m=\sigma_{s_{m-2}}\sum_{k=1}^{L_m}\frac{k-1}{L_m}\Delta^n_{N_{m-1}+k}W,\ \beta'_m=\sigma_{s_{m-2}}\sum_{k=1}^{L_{m+1}}\Big(1-\frac{k-1}{L_{m+1}}\Big)\Delta^n_{N_m+k}W,\end{cases}$$

then $E\Big[\alpha^2_{m-1}|\mathcal{G}_{m-1}\Big]=\sigma^2_{s_{m-2}}w_{m,1}, E\Big[\beta^2_m\Big|\mathcal{G}_{m-1}\Big]=\sigma^2_{s_{m-2}}w_{m,2}$ and $E\Big[\Big(\alpha'_{m-1}\Big)^2\Big|\mathcal{G}_{m-1}\Big]=\sigma^2_{s_{m-2}}w_{m+1,1}, E[(\beta'_m)^2\Big|\mathcal{G}_{m-1}\Big]=\sigma^2_{s_{m-2}}w_{m+1,2}$ and conditional on $\mathcal{G}_{m-1}$, $\alpha_m,\beta_m,\beta'_m$ are mutually independent, $\alpha_m,\alpha'_m,\beta'_m$ are mutually independent, but $\beta_m$ and $\alpha'_m$ are correlated.

Therefore, we have

$$\begin{aligned}&E\Big[\Big(\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\Big)^2\Big(\bar{X}_{s_{m+1}}-\bar{X}_{s_m}\Big)^2\Big|\mathcal{G}_{m-1}\Big]\\&=E\Big[\Big(\alpha_m+\beta_m)^2\Big(\alpha'_m+\beta'_m\Big)^2\Big|\mathcal{G}_{m-1}\Big]+o_P\Big(\Delta_M^2\Big)\\&=E\Big[\Big(\alpha_m^2+\beta_m^2+2\alpha_m\beta_m\Big)\Big(\Big(\alpha'_m\Big)^2+\Big(\beta'_m\Big)^2+2\alpha'_m\beta'_m\Big)\Big|\mathcal{G}_{m-1}\Big]+o_P\Big(\Delta_M^2\Big)\\&=\sigma^4_{s_{m-2}}\Big(w_{m,1}w_{m+1,1}+w_{m,1}w_{m+1,2}+\tilde{v}_{m,1}+w_{m,2}w_{m+1,2}\Big)\Delta_M^2\\&\quad+o_P\Big(\Delta_M^2\Big),\end{aligned} \tag{54}$$

where, $\tilde{v}_{m,1}:=\sigma^4_{s_{m-2}}E\Big[\sum_{k=1}^{L_m}\Big(1-\frac{k-1}{L_m}\Big)\Delta^n_{N_{m-1}+k}W\Big]^2\Big[\sum_{k=1}^{L_m}\Big(\frac{k-1}{L_m}\Big)\Delta^n_{N_{m-1}+k}W\Big]^2$. Then, the two dimension Itô's formula is applied to obtain

$$\tilde{v}_{m,1}=\frac{1}{L_m^6}\Big\{\sum_{k=1}^{L_m-1}\sum_{j=0}^{k}\Big(k^2(L_m-j)^2+(L_m-k)^2j^2+4kj(L_m-k)(L_m-j)\Big)\Big\}.$$

Combining (52)–(54) and

$$w_{m,1}w_{m+1,1}+w_{m,1}w_{m+1,2}+w_{m,2}w_{m+1,2}-w_mw_{m+1}=-w_{m,2}w_{m+1,1}, \tag{55}$$

simple computation yields

$$\tilde{v}_{m,2}:=-w_{m,2}w_{m+1,1}=-\frac{1}{L_m^6}\sum_{k=1}^{L_m-1}\sum_{j=0}^{L_m-1}\Big(L_m-j\Big)^2k^2. \tag{56}$$

Finally, by considering (49), we give the asymptotic variance in (13).

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

**Proof of Theorem 2.** Here, we present the proof for the case of $L_i \equiv L$, the proof for the general case is similar. Denote

$$\xi_m := \sqrt{M}\sum_{k=1}^{2L} g_2\Big(\frac{k-1}{2L}\Big)\Delta^n_{(m-2)L+k}X,$$
$$\beta_m := \sqrt{M}\sigma_{s_{m-2}}\sum_{k=1}^{2L} g_2\Big(\frac{k-1}{2L}\Big)\Delta^n_{(m-2)L+k}W. \tag{57}$$

We have

$$M^{p/2-1}\sum_{m=1}^{M}\Big|\frac{1}{L}\sum_{j=1}^{L}X_{t_{\left(m-1\right)L+j}} - \frac{1}{L}\sum_{j=1}^{L}X_{t_{\left(m-2\right)L+j}}\Big|^p = \frac{1}{M}\sum_{m=1}^{M}\Big|\xi_m\Big|^p. \tag{58}$$

Observe that

$$\sum_{m=1}^{M}E\Big[\Big(\frac{1}{M}\Big|\beta_m\Big|^p\Big)^2\Big|\mathcal{G}_{m-1}\Big] = \frac{1}{M^2}\sum_{m=1}^{M}m_{2p}\Big|\sigma_{s_{m-2}}\Big|^{2p}w^p \xrightarrow{P} 0, \tag{59}$$

as $M \to \infty$ by the boundedness of $\sigma$. Because $\beta_m$ is $\mathcal{G}_{m+1}$-measurable, by Lemma 5.2 of Barndorff-Nielsen et al. (2006), we have

$$\frac{1}{M}\sum_{m=1}^{M}\Big[\Big|\beta_m\Big|^p - E\Big[\Big|\beta_m\Big|^p\Big|\mathcal{G}_{m-1}\Big]\Big] \xrightarrow{P} 0. \tag{60}$$

Riemann integrability yields

$$\frac{1}{M}\sum_{m=1}^{M}E\Big[\Big|\beta_m\Big|^p\Big|\mathcal{G}_{m-1}\Big] = \frac{1}{M}\sum_{m=1}^{M}m_p\Big|\sigma_{s_{m-2}}\Big|^p\Big(w\Big)^{p/2} \to m_p(w)^{p/2}\int_0^t\sigma_s^p ds \tag{61}$$

point-wise in $\omega$. Thus, to prove the theorem, it suffices to show that

$$\frac{1}{M}\sum_{m=1}^{M}\Big(\Big|\xi_m\Big|^p - \Big|\beta_m\Big|^p\Big) \xrightarrow{P} 0. \tag{62}$$

We utilize Hölder's and Burkholder's inequalities and an elementary inequality to obtain the following.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

- $p \geq 2$, $\left|\left|x+y\right|^p - \left|x\right|^p\right| \leq C\left(\left|y\right|^p + \left|x\right|^{p-1}\left|y\right|\right)$:

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M} E\left|\left|\xi_m\right|^p - \left|\beta_m\right|^p\right| &\leq \frac{C}{M}\sum_{m=1}^{M} E\left|\xi_m-\beta_m\right|^p + \left|\xi_m-\beta_m\right|\left|\beta_m\right|^{p-1}) \\
&\leq \frac{C}{M}\Big(\sum_{m=1}^{M} E\Big|\sqrt{M}\int_{s_{m-2}}^{s_m} g_2\Big(\frac{\lfloor (s-s_{m-2})/(2\Delta_M)\rfloor}{2L}\Big)\Big(\sigma_s-\sigma_{s_{m-2}}\Big)dW_s\Big|^p \\
&\quad + \sum_{m=1}^{M}\sqrt{E|\xi_m-\beta_m|^2 E|\beta_m|^{2p-2}}\Big) \\
&\leq \frac{C}{M}\sum_{m=1}^{M} E\Big|M\int_{s_{m-2}}^{s_m} g_2\Big(\frac{\lfloor (s-s_{m-2})/(2\Delta_M)\rfloor}{2L}\Big)\Big(\sigma_s-\sigma_{s_{m-2}}\Big)^2 ds\Big|^{p/2} \\
&\quad + C\Delta_M^{1/2} \\
&\leq C\sum_{m=1}^{M}\int_{s_{m-2}}^{s_m} g_2^{p/2}\Big(\frac{\lfloor (s-s_{m-2})/(2\Delta_M)\rfloor}{2L}\Big) E\Big(\sigma_s-\sigma_{s_{m-2}}\Big)^p ds + C\Delta_M^{1/2} \\
&\leq C\Delta_M^{p/2} + C\Delta_M^{1/2} \to 0.
\end{aligned}$$

Combining this with (58)–(61), we have

$$M^{\frac{p}{2}-1}\sum_{m=1}^{M}\left|\bar{X}_{s_m}-\bar{X}_{s_{m-1}}\right|^p \xrightarrow{P} m_p(w)^{p/2}\int_0^t \sigma_s^p ds. \qquad \blacksquare$$

**Proof of Theorem 3.** In view of the stable convergence, the proof follows from Theorems 1 and 2. ■

**Proof of Theorem 4.** (1) **Consistency** The required result follows from the following results

$$\sum_{m=1}^{M}\Big(X_{R_m}-X_{R_{m-1}}\Big)^2 - \sum_{m=1}^{M}\sigma_{R_{m-1}}^2\Big(W_{R_m}-W_{R_{m-1}}\Big)^2 \xrightarrow{P} 0, \tag{63}$$

$$\sum_{m=1}^{M}\sigma_{R_{m-1}}^2\Big(W_{R_m}-W_{R_{m-1}}\Big)^2 - \sum_{m=1}^{M}\sigma_{R_{m-1}}^2\Big(R_m-R_{m-1}\Big) \xrightarrow{P} 0, \tag{64}$$

$$\sum_{m=1}^{M}\sigma_{R_{m-1}}^2\Big(R_m-R_{m-1}\Big) \xrightarrow{P} \int_0^t \sigma_s^2 ds. \tag{65}$$

Defining a new filtration $\mathcal{G}'_m := \mathcal{G}_m \vee \sigma\{U_i, i=1,2,\cdots,M\}$, (63), (64), and (65) can be shown similarly as the deterministic sampling case by dominated convergence theorem.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

**Proof of Theorem 4.** (2) **Central limit theorem** We prove the theorem by showing the following:

$$\sqrt{M}\Big(\sum_{m=1}^{M}\Big(X_{R_m}-X_{R_{m-1}}\Big)^2-\sum_{m=1}^{M}\sigma^2_{s_{m-2}}\Big(W_{R_m}-W_{R_{m-1}}\Big)^2\Big)\xrightarrow{P}0, \tag{66}$$

$$\begin{aligned}&\sqrt{M}\sum_{m=1}^{M}\Big(\sigma^2_{s_{m-2}}\Big(W_{R_m}-W_{R_{m-1}}\Big)^2-\sigma^2_{s_{m-2}}\int_{R_{m-1}}^{R_m}ds\Big)\\&=:\sum_{i=1}^{M}\varrho_m\xrightarrow{S}\sqrt{2}\int_0^t(Q'_u)^{1/2}\sigma_s^2dW''_u,\end{aligned} \tag{67}$$

$$\sqrt{M}\sum_{m=1}^{M}\Big(\sigma^2_{s_{m-2}}\int_{R_{m-1}}^{R_m}ds-\int_{R_{m-1}}^{R_m}\sigma_s^2ds\Big)\xrightarrow{P}0, \tag{68}$$

$$\sqrt{M}\int_0^{R_1}\sigma_s^2ds\xrightarrow{P}0,\quad \sqrt{M}\int_{R_M}^{t}\sigma_s^2ds\xrightarrow{P}0. \tag{69}$$

(69) is trivial. (66) and (68) can be shown by repeating the methods employed in the proof of previous theorems. Let $\mathcal{H}_m=\mathcal{G}_m\vee\sigma\{U_i,1\le i\le m-1\}$. We have $E\Big[\varrho_m|\mathcal{H}_{m-1}\Big]=0$ and $E\Big[\varrho_m^4|\mathcal{H}_{m-1}\Big]=o_P(\Delta_M)$. From (24), we have

$$E\Big[\varrho_m^2|\mathcal{H}_{m-1}\Big]=2\sqrt{M}\sigma^4_{s_{m-2}}E\Big[(R_m-R_{m-1})^2|\mathcal{H}_{m-1}\Big]=2\chi_m\Delta_M.$$

In addition, considering the orthogonality discussed above and defining $Q(u)=\lim_{M\to\infty}\sum_{s_m\le u}2\chi_m\Delta_M$, we obtain (67).

**Proof of Theorem 4.** (3) **Studentized central limit theorem** The proof follows from Theorem 4 (CLT), by employing (27) and stable convergence. ■

## NOTE

1. See Figure 1 for a graphical explanation.

## REFERENCES

Aït-Sahalia, Y., J. Fan, & D. Xiu (2010) High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association* 105(492), 1504–1517.

Aït-Sahalia, Y. & J. Jacod (2009) Estimating the degree of activity of jumps in high frequency financial data. *Annals of Statistics* 37(5A), 2202–2244.

Aït-Sahalia, Y. & P. Mykland (2003) The effects of random and discrete sampling when estimating continuous-time diffusions. *Econometrica* 71(2), 483–549.

Aït-Sahalia, Y., P. Mykland, & L. Zhang (2011) Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics* 160(1), 160–175.

Aït-Sahalia, Y. & P.A. Mykland (2004) Estimators of diffusions with randomly spaced discrete observations: A general theory. *Annals of Statistics* 32(5), 2186–2222.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

Barndorff-Nielsen, O., S. Graversen, J. Jacod, M. Podolskij, & N. Shephard (2006) A central limit theorem for realised power and bipower variations of continuous semimartingales. In Yu. Kabanov, R. Liptser, J. Stoyanov (eds.), *From Stochastic Analysis to Mathematical Finance; the Shiryaev Festschrift*. Springer.

Barndorff-Nielsen, O.E., P.R. Hansen, A. Lunde, & N. Shephard (2008a) Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica* 76(6), 1481–1536.

Barndorff-Nielsen, O.E., P.R. Hansen, A. Lunde, & N. Shephard (2008b) Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics* 162(2), 149–169.

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

Barndorff-Nielsen, O.E. & N. Shephard (2006) Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics* 4(1), 1–30.

Hayashi, T., J. Jacod, & N. Yoshida (2011) Irregular sampling and central limit theorems for power variations: The continuous case. *Annales de l'Institut Henri Poincare Probability and Statistics* 47(4), 1197–1218.

Jacod, J. (2012) Statistics and high frequency data. In M. Kessler, A. Lindner, & M. S$\phi$rensen (eds.) *Proceedings of the 7th Séminaire Européen de Statistique, La Manga, 2007: Statistical Methods for Stochastic Differential Equations*.

Jacod, J., Y. Li, P.A. Mykland, M. Podolskij, & M. Vetter (2009) Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Processes and Their Applications* 119(7), 2249–2276.

Jacod, J., M. Podolskij, & M. Vetter (2010) Limit theorems for moving averages of discretized processes plus noise. *Annals of Statistics* 38(3), 1478–1545.

Jacod, J. & A.V. Shiryayev (2003) *Limit Theorems for Stochastic Processes*. Springer.

Jing, B.Y., X.B. Kong, Z. Liu, & P.A. Mykland (2012) On the jump activity index for semimartingales. *Journal of Econometrics* 166(2), 213–223.

Li, Y., P. Mykland, E. Renault, L. Zhang, & X. Zheng (2014) Realized volatility when sampling time are possibly endogenous. *Econometric Theory* 30, 580–605.

Mancini, C. (2009) Nonparametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics* 36, 270–296.

Mörters, P. & Y. Peres (2010) *Brownian Motion*. Cambridge University Press.

Mykland, P. & L. Zhang (2009) Inference for continuous semimartingales observed at high frequency: A general approach. *Econometrica* 77(5), 1403–1445.

Mykland, P.A. & L. Zhang (2006) ANOVA for diffusion and Itô processes. *Annals of Statistics* 34(4), 1931–1963.

Mykland, P.A. & L. Zhang (2012) The econometrics of high frequency data. In M. Kessler, A. Lindner and M. Sorensen, (eds.), *Statistical Methods for Stochastic Differential Equations*, 109–185.

Podolskij, M. & M. Vetter (2009) Bipower-type estimation in a noisy diffusion setting. *Stochastic Processes and Their Applications* 119(9), 2803–2831.

Renyi, A. (1963) On stable sequences of events. *Sankhya: The Indian Journal of Statistics, Series A* 25(3), 293–302.

Todorov, V. & T. Bollerslev (2010) Jumps and betas: A new framework for disentangling and estimating systematic risks. *Journal of Econometrics* 157, 220–235.

Todorov, V. & G. Tauchen (2010) Activity signature functions for high frequency data analysis. *Journal of Econometrics* 154, 125–138.

Tsay, R.S. (2005) *Analysis of Financial Time Series*. Wiley.

Xiu, D. (2010) Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics* 159, 235–250.

Zhang, L. (2006) Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli* 12(6), 1019–1043.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press

Zhang, L. (2011) Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics* 160(1), 33–47.

Zhang, L., P. Mykland, & Y. Aït-Sahalia (2005) A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100(472), 1394–1411.

https://doi.org/10.1017/S0266466615000420 Published online by Cambridge University Press