Detecting factors of quadratic variation in the presence of market microstructure noise

Naoto Kunitomo¹ · Daisuke Kurisu²

Received: 25 August 2020 / Accepted: 22 December 2020 / Published online: 4 February 2021
© The Author(s) 2021

Abstract

A method of detecting latent factors of quadratic variation (QV) of Itô semimartingales from a set of discrete observations is developed when the market microstructure noise is present. We propose a new way to determine the number of latent factors of quadratic co-variations of asset prices based on the SIML (separating information maximum likelihood) method by Kunitomo et al. (Separating information maximum likelihood estimation for high frequency financial data. Springer, Berlin, 2018). In high-frequency financial data, it is important to investigate the effects of possible jumps and market microstructure noise existed in financial markets. We explore the estimated variance–covariance matrix of latent (efficient) prices of the underlying Itô semimartingales and investigate its characteristic roots and vectors of the estimated quadratic variation. We give some simulation results to see the finite sample properties of the proposed method and illustrate an empirical data analysis on the Tokyo stock market.

Keywords Itô-semimartingales · High-frequency financial data · Market microstructure noise · Quadratic variation · Hidden factors · SIML estimation · Characteristic roots and vectors · Limiting distributions

Mathematics Subject Classification MSC62P20 · 93E03

✉ Naoto Kunitomo
naoto.kunitomo@gmail.com

Daisuke Kurisu
kurisu.d.aa@m.titech.ac.jp

¹ School of Political Science and Economics, Meiji University, Sarugakucho 3rd Building C-106, 1-1 Kanda-Surugadai, Chiyoda-ku, Tokyo 101-8301, Japan

² School of Engineering, Tokyo Institute of Technology, Meguro-ku, Ōokayama 2-12-1 W9-96, Tokyo 152-8552, Japan

1 Introduction

In financial econometrics, several statistical methods have been proposed to estimate the integrated volatility and co-volatility from high-frequency data. The integrated volatility is one type of Lévy or Brownian functionals, and the realized volatility (RV) estimate has been often used when there does not exist any market microstructure noise and the underlying continuous time process is directly observed. It has been known that the RV estimator is quite sensitive to the presence of market microstructure noise in high-frequency financial data. Then, several statistical methods have been proposed to estimate the integrated volatility and co-volatility. See Aït-Sahalia and Jacod (2014) for the detail of recent developments of financial econometrics. In particular, Malliavin and Mancino (2002, 2009) have developed the Fourier series method, which is related to the SIML (separating information maximum likelihood) estimation by Kunitomo et al. (2018) used in this paper. See Gloter and Jacod (2001), Bibinger et al. (2014), Christensen et al. (2010), and Mancino et al. (2017) for related topics.

In this paper, we develop a new statistical way of detecting latent factors of quadratic variation of Itô-semimartingales from a set of discrete observations when the market microstructure noise is present. We will use the high-frequency asymptotic method such that the length of observation intervals becomes small as the number of observations grows, which has been often used in recent financial econometrics. In finance, it is important to find several latent factors among many financial prices such as stocks, bonds and other financial products. It might be a practice to find latent factors after calculating various returns from price data and apply statistical tools such as principal component analysis, factor analysis and other statistical multivariate techniques. However, it should be noted that the standard statistical analysis has been developed to analyze independent (or stationary) observations and most financial prices are classified neither as independent nor stationary observations. In addition to this fact, it is important to notice that when we have market microstructure noises or measurement errors for prices, we have another statistical problem when we use high-frequency financial data. Although the multivariate statistical analyses such as principal components and factor models have been applied to financial data, these statistical methods do not necessarily give the right answers when we have market microstructure noise in high-frequency data. The standard statistical procedures could be a misleading way to analyze high-frequency financial data.

To develop a way to determine the number of latent factors of quadratic covariation or the integrated volatility of asset prices, we shall utilize the SIML (separating information maximum likelihood) method, which was originally developed by Kunitomo and Sato (2013, 2021) and Kunitomo et al. (2018). In high-frequency financial data, it is important to investigate the effects of the possible jumps and the market microstructure noise existed in financial markets. We explore the estimation problem of the variance-covariance matrix of the underlying Itô semimartingales, that is, the quadratic variation (QV). We shall show that it is possible to derive the asymptotic properties of the characteristic vectors

and roots of the estimated QV and then develop some test statistic for the rank condition. Our estimators of characteristic vectors and roots are consistent and have desirable asymptotic properties. We develop some test statistics based on the characteristic roots and vectors to detect the number of factors of QV. We also give a real data analysis on the Tokyo stock market as an illustration.

There exist some methods for the analysis of volatility structure under high-frequency observations [e.g., Aït-Sahalia and Xiu (2019), Fissler and Podolskij (2017) and Jacod and Podolskij (2013)]. However, these papers aim at detecting volatility structures that are different from our formulation to be investigated in the present study. Aït-Sahalia and Xiu (2019) investigated the estimation problem of number of factors in high-dimensional factor models which are related to our setting, but their estimation method is different from ours. Jacod and Podolskij (2013) developed a method to detect the “maximal” rank of the volatility process with the fixed time interval framework. Fissler and Podolskij (2017) extended their method to noisy high-frequency data. Since our goal is to detect the rank of “integrated” volatility (or quadratic variation) directly, the main purpose of these studies is different from ours. Our method is simple, and it is not difficult to be implemented even when the dimension is not small and there are some jump terms as well as microstructure noise.

In Sect. 2, we define the Itô semimartingale and quadratic variation, which is an extension of the integrated volatility with jump parts. Then, we define the SIML estimation and its asymptotic property for Itô semimartingales. In Sect. 3, we consider the characteristic equation of the estimated (conditional) variance-covariance matrix and give the theoretical results on the asymptotic properties of the associated characteristic roots and vectors when the true (or efficient price) process is an Itô semimartingale and there are market microstructure noises. Then, we propose test statistics for the rank condition of the quadratic variation, which can be applied to detect the number of latent factors of integrated volatilities for continuous diffusion processes as special cases. In Sect. 4, we give some results on the Monte Carlo simulations of our procedure, and in Sect. 5, we illustrate an empirical data analysis on the Tokyo stock market. Then, in Sect. 6, we give concluding remarks. Some mathematical details are given in Appendix.

2 Estimation of quadratic variation under market microstructure noise

We consider a continuous-time financial market in a fixed terminal time T , and we set $T = 1$ without loss of generality. The underlying log-price is a p -dimensional Itô semimartingale, but we focus on the fact that we observe the log-price process in high-frequency financial prices and they are contaminated by the market microstructure noise. We define the filtered probability space on which the prices follow the Itô semimartingale in the presence of market microstructure noise.

Let the first filtered probability space be $(\Omega^{(0)}, \mathcal{F}^{(0)}, (\mathcal{F}_t^{(0)})_{t \geq 0}, P^{(0)})$ on which the p -dimensional Itô semimartingale $\mathbf{X} = (\mathbf{X}(t))_{0 \leq t \leq 1}$ is defined. We adopt the

construction of the whole filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [0,1]}, P)$, where both the process $\mathbf{X}$ and the noise are defined.

We set $(\Omega^{(1)}, \mathcal{F}^{(1)}, (\mathcal{F}_t^{(1)})_{t \in [0,1]}, P^{(1)})$ as the second filtered probability space, where $\Omega^{(1)}$ is the total events of micromarket noise, $\mathcal{F}^{(1)}$ is the Borel σ -field on $\Omega^{(1)}$, and $P^{(1)}$ is the probability measure. The market microstructure noise process $\mathbf{v} = (\mathbf{v}(t))_{t \in [0,1]}$ as the process on $(\Omega^{(1)}, \mathcal{F}^{(1)}, (\mathcal{F}_t^{(1)})_{t \in [0,1]}, P^{(1)})$ with the filtration $\mathcal{F}_t^{(1)} = \sigma(\mathbf{v}(s) : s \leq t)$ for $0 \leq t \leq 1$. We use the filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [0,1]}, P)$, where $\Omega = \Omega^{(0)} \times \Omega^{(1)}$, $\mathcal{F} = \mathcal{F}^{(0)} \times \mathcal{F}^{(1)}$, $\mathcal{F}_t = \mathcal{F}_t^{(0)} \otimes \mathcal{F}_t^{(1)}$ and $P = P^{(0)} \times P^{(1)}$.

When we consider the continuous time stochastic processes, the class of Itô semimartingales is a fundamental one and it includes the diffusion processes and jump processes as special cases. In applications of high-frequency financial data, it has been known that the role of market microstructure noise is important. However, it is not straightforward to estimate the volatility and co-volatilities or quadratic variation in the general case in the presence of market microstructure noise.

2.1 Itô semimartingale and quadratic variation

In this section, we describe the statistical model of the present paper. Let $\mathbf{Y}(t_i^n) = (Y_g(t_i^n))_{g=1, \dots, p}$ be the (p -dimensional) observed (log-)prices at $t_i \in [0, 1]$ and $i = 1, \dots, n$, which satisfies

$$\mathbf{Y}(t_i^n) = \mathbf{X}(t_i^n) + \mathbf{v}(t_i^n) \quad (i = 1, \dots, n), \quad (1)$$

where $\mathbf{X}(t_i^n) = (X_g(t_i^n))$ is the $p \times 1$ hidden stochastic vector process and $\mathbf{v}(t_i^n) = (v_g(t_i^n))$ is a sequence of (mutually) independently and identically distributed market microstructure noises with $\mathcal{E}[\mathbf{v}(t_i^n)] = \mathbf{0}$ and $\mathcal{E}[\mathbf{v}(t_i^n)\mathbf{v}(t_i^n)'] = \Sigma_v (> 0$ a positive definite matrix). We set the fixed initial condition $\mathbf{X}(t_0^n)$, $\mathbf{Y}(t_0) = \mathbf{X}(t_0^n)$, and we take $t_i^n - t_{i-1}^n = 1/n$ (an equidistance interval) and consider the situation $n \rightarrow \infty$.

We assume that these market microstructure noises are independent of the p -dimensional continuous-time stochastic process $\mathbf{X}(t)$, which follows

$$\begin{aligned} \mathbf{X}(t) = & \mathbf{X}(0) + \int_0^t \mathbf{b}(s) ds + \int_0^t \sigma(s) d\mathbf{W}(s) + \int_0^t \int_{\|\mathbf{u}\| < 1} \Delta(s, \mathbf{u})(\boldsymbol{\mu} - \mathbf{v})(ds, d\mathbf{u}) \\ & + \int_0^t \int_{\|\mathbf{u}\| \geq 1} \Delta(s, \mathbf{u}) \boldsymbol{\mu}(ds, d\mathbf{u}), \end{aligned} \quad (2)$$

where $\mathbf{b}(s)$ and $\sigma(s)$ are the p -dimensional adapted drift process and the $p \times q_1$ ($1 \leq q_1 \leq p$) instantaneous predictable volatility process, $\mathbf{W}(s) = (W_g(s))$ is the $q_1 \times 1$ standard Brownian motions, $\Delta(\omega, s, \mathbf{u})$ is a $\mathbf{R}^p$ -valued predictable function on $\Omega \times [0, \infty) \times \mathbf{R}^{q_2}$ ($q_2 \leq p$), $\boldsymbol{\mu}(\cdot)$ is a Poisson random measure on $[0, \infty) \times \mathbf{R}^{q_2}$, which is independent of $\mathbf{W}(s)$, and $\mathbf{v}(ds, d\mathbf{u}) = ds \otimes \lambda(d\mathbf{u})$ is the predictable compensator or intensity measure of $\boldsymbol{\mu}$ with a σ -finite measure λ on $(\mathbf{R}^p, \mathcal{B}^p)$. We partially follow the notation used in Jacod and Protter (2012). The jump terms are denoted as $\Delta \mathbf{X}(s) = (\Delta X_g(s))$ ($\Delta X_g(s) = X_g(s) - X_g(s-)$, $X_g(s-) = \lim_{u \uparrow s} X_g(u)$ at any $s \in [0, 1]$), and $\|\cdot\|$ is the Euclidean norm on $\mathbf{R}^p$. We use the notation

$\mathbf{c}(s) = \boldsymbol{\sigma}(s)\boldsymbol{\sigma}'(s) = (\mathbf{c}_{gh}(s))$ ($p \times p$ matrix) and for $p \times 1$ vectors $\mathbf{y}_i = \mathbf{Y}(t_i^n)$, $\mathbf{x}_i = \mathbf{X}(t_i^n)$, and $\mathbf{v}_i = \mathbf{v}(t_i^n)$ ($i = 1, \dots, n$).

In this paper, we restrict our formulation in (1) and (2) although there can be some extensions. We consider the volatility function $\boldsymbol{\sigma}(s) = (\sigma_{gh}(s))$, $g = 1, \dots, p; h = 1, \dots, q_1$ is deterministic or its elements follow Itô's Brownian continuous semi-martingale given by

$$\sigma_{gh}^{(x)}(t) = \sigma_{gh}^{(x)}(0) + \int_0^t \mu_{gh}^{\sigma}(s)ds + \int_0^t \omega_{gh}^{\sigma}(s)d\mathbf{W}^{\sigma}(s), \quad (3)$$

where $\mathbf{W}^{\sigma}(s)$ is a $q_3 \times 1$ ($q_3 \geq 0$) Brownian motions, which can be correlated with $\mathbf{W}(s)$, and $\mu_{gh}^{\sigma}(s)$ are the drift terms of volatilities, and $\omega_{gh}^{\sigma}(s)$ ($1 \times q_3$) are the diffusion terms of instantaneous volatilities, respectively. They are Hölder-continuous, predictable and progressively measurable with respect to $(\boldsymbol{\Omega}^{(0)}, \mathcal{F}^{(0)}, (\mathcal{F}_t^{(0)})_{t \geq 0}, P^{(0)})$. We consider the case when they are bounded and Hölder-continuous such that the volatility and co-volatility processes are smooth.

For the resulting simplicity, we set several assumptions on (1)–(3), which can be certainly relaxed to some extent.

Assumption 1

- (a) The drift function $\mathbf{b}(\omega, t) = \mathbf{0}$ for $0 \leq t \leq 1$.
- (b) The elements of volatility matrix are bounded and the process $\boldsymbol{\sigma}$ is Hölder-continuous. We have $\int_t^{t+u} \|\boldsymbol{\sigma}(s)\|ds > 0$ a.s. for all $t, u > 0$.
- (c) The jump coefficients $\boldsymbol{\Delta}(\omega, t, \mathbf{u})$ are bounded and deterministic.
- (d) The noise terms $\mathbf{v}(t_i^n) (= \mathbf{v}_g(t_i^n))$ ($i = 1, \dots, n; g = 1, \dots, p$) are a sequence of i.i.d. random variables with $\mathcal{E}[\mathbf{v}(t_i^n)] = \mathbf{0}$, $\mathcal{E}[\mathbf{v}(t_i^n)\mathbf{v}'(t_i^n)] = \boldsymbol{\Sigma}_v$ (a positive definite matrix) and $\mathcal{E}[v_g^4(t_i^n)] < +\infty$ ($g = 1, \dots, p$). Furthermore, the stochastic processes of $\mathbf{v}$ and $\mathbf{X}$ are independent.
- (e) We have equidistance observations, that is, $t_0^n = 0$ and $t_i^n = i/n$ ($i = 1, \dots, n$) in $[0, 1]$.

Conditions (a)–(c) are stronger than the ones in some literature, but they are sufficient for the existence of solutions for the stochastic differential equations (SDE). (See Chapter IV of Ikeda and Watanabe 1989, for instance.) We consider only the cases under conditions (d) and (e) in the following analysis. However, it is possible to extend the following analysis to more general cases. For instance, it may be straightforward to consider the case when the volatility functions have jump terms with some complication of our analysis. Since this paper is the first attempt to develop a new way to handle the problem, we shall consider the simple cases in this paper.

The fundamental quantity for the continuous-time Itô semimartingale with $p \geq 1$ is the quadratic variation (QV) matrix, which is given by

$$\Sigma_x = \int_0^1 c(s)ds + \sum_{0 \leq s \leq 1} (\Delta \mathbf{X}(s))(\Delta \mathbf{X}(s))' = (\Sigma_{gh}^{(x)}). \quad (4)$$

When the stochastic process is the diffusion-type, Σ_x becomes the integrated volatility $\int_0^1 c(s)ds$. The class of Itô semimartingales and the quadratic variation, which have been standard in stochastic analysis, are fully explained by Ikeda and Watanabe (1989) and Jacod and Protter (2012) as the standard literature.

2.2 On the SIML estimation

Kunitomo and Sato (2013) have developed the separating information maximum likelihood (SIML) estimation for general $p \geq 1$, but there are no jump terms. The SIML estimator of $\hat{\Sigma}_x$ for the integrated volatility is defined by

$$\hat{\Sigma}_x = \mathbf{G}_m = \frac{1}{m_n} \sum_{k=1}^{m_n} \mathbf{z}_k \mathbf{z}_k' = (\hat{\Sigma}_{gh}^{(x)}), \quad (5)$$

where $\mathbf{z}_k = (z_{jk})$ ($j = 1, \dots, p; k = 1, \dots, m_n$), which are constructed by the transformation from $\tilde{\mathbf{Y}}_n = (\mathbf{y}_i')$ ($n \times p$) to $\mathbf{Z}_n = (\mathbf{z}_k')$ by

$$\mathbf{Z}_n = \mathbf{K}_n(\mathbf{Y}_n - \tilde{\mathbf{Y}}_0) \quad (6)$$

where $\mathbf{K}_n = h_n^{-1/2} \mathbf{P}_n \mathbf{C}_n^{-1}$, $h_n = 1/n$,

$$\mathbf{C}_n^{-1} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}_{n \times n},$$

$$\mathbf{P}_n = (p_{jk}^{(n)}), \quad p_{jk}^{(n)} = \sqrt{\frac{2}{n + \frac{1}{2}}} \cos \left[\frac{2\pi}{2n+1} \left(k - \frac{1}{2} \right) \left(j - \frac{1}{2} \right) \right]$$

and $\tilde{\mathbf{Y}}_0 = \mathbf{1}_n \cdot \mathbf{y}_0'$.

By using the spectral decomposition $\mathbf{C}_n^{-1} \mathbf{C}_n'^{-1} = \mathbf{P}_n \mathbf{D}_n \mathbf{P}_n'$ and $\mathbf{D}_n$ is a diagonal matrix with the k th element $d_k = 2[1 - \cos(\pi(\frac{2k-1}{2n+1}))]$ ($k = 1, \dots, n$) and $a_{kn} (= n \times d_k) = 4n \sin^2 \left[\frac{\pi}{2} \left(\frac{2k-1}{2n+1} \right) \right]$.

To assure some desirable asymptotic properties of the SIML estimator, we need the condition that the number of terms m_n should be dependent on n and we need the order requirement that $m_n = O(n^\alpha)$ ($0 < \alpha < 0.5$) for the consistency and $m_n = O(n^\alpha)$ ($0 < \alpha < 0.4$) for the asymptotic normality.

The variance-covariance matrix Σ_v can be consistently estimated by

$$\hat{\Sigma}_v = \frac{1}{l_n} \sum_{k=n+1-l_n}^n a_{kn}^{-1} \mathbf{z}_k \mathbf{z}_k', \quad (7)$$

where $l_n = [n^\beta]$ ($0 < \alpha < \beta < 1$). We can take β being slightly less than 1 and $\hat{\Sigma}_v = \Sigma_v + O_p(\frac{1}{\sqrt{l_n}})$ such that the effects of estimating Σ_v are negligible. (See Chapter 5 of Kunitomo et al. 2018.)

When $\mathbf{X}$ is an Itô semimartingale with possible jumps, the asymptotic properties of the SIML estimator were stated in Chapter 9 of Kunitomo et al. (2018) (Proposition 9.1 and Corollary 9.1) without the detailed exposition. Because they are the starting points of further developments, we state an extended version of their result and we give some supplementary derivations in Appendix for the sake of convenience.

In the following results, we use the stable convergence arguments and $\mathcal{F}^{(0)}$ -conditionally Gaussianity, which have been developed and explained by Jacod (2008) and Jacod and Protter (2012), and use the notation $\xrightarrow{\mathcal{L}-s}$ as stable convergence in law. For the general reference on stable convergence, we refer to Häusler and Luschgy (2015). We use the notation $\xrightarrow{d}$ and $\xrightarrow{p}$ as convergence in distribution and in probability, respectively.

Theorem 2.1 *Suppose Assumption 1 is satisfied and $\mathcal{E}[v_i^4(t_i^n)] < +\infty$ ($g = 1, \dots, p$) in (1), (2) and (3).*

- (i) For $m_n = [n^\alpha]$ ($[\cdot]$ is the floor function) and $0 < \alpha < 0.5$, as $n \rightarrow \infty$

$$\hat{\Sigma}_x - \Sigma_x \xrightarrow{p} \mathbf{O}. \quad (8)$$

- (ii) Let

$$\hat{\Sigma}_{gh}^{(*)} = \sqrt{m_n} \left[\hat{\Sigma}_{gh}^{(x)} - \Sigma_{gh}^{(x)} \right], \quad (9)$$

where $\hat{\Sigma}_{gh}^{(x)}$ is the (g, h) th component of $\hat{\Sigma}_x$ and $\Sigma_{gh}^{(x)}$ is the (g, h) th component of Σ_x . Then, as $n \rightarrow \infty$, for $m_n = [n^\alpha]$ and $0 < \alpha < 0.4$, we have that

$$\begin{bmatrix} \hat{\Sigma}_{gh}^{(*)} \\ \hat{\Sigma}_{kl}^{(*)} \end{bmatrix} \xrightarrow{\mathcal{L}-s} N \left[\mathbf{0}, \begin{pmatrix} V_{gh} & V_{gh,kl} \\ V_{gh,kl} & V_{kl} \end{pmatrix} \right] \quad (10)$$

where $\mathbf{c}(s) = (c_{gh}(s))$ ($0 \leq s \leq 1$),

$$\begin{aligned} V_{gh} = & \int_0^1 \left[\mathbf{c}_{gg}(s) \mathbf{c}_{hh}(s) + \mathbf{c}_{gh}^2(s) \right] ds \\ & + \sum_{0 < s \leq 1} \left[\mathbf{c}_{gg}(s) (\Delta X_h(s))^2 + \mathbf{c}_{hh}(s) (\Delta X_g(s))^2 + 2\mathbf{c}_{gh}(s) (\Delta X_g(s) \Delta X_h(s)) \right] \end{aligned}$$

and

$$\begin{aligned}
V_{gh,kl} = & \int_0^1 [\mathbf{c}_{gk}(s)\mathbf{c}_{hl}(s) + \mathbf{c}_{gl}(s)\mathbf{c}_{hk}(s)] ds \\
& + \sum_{0 < s \leq 1} [\mathbf{c}_{gk}(s)\Delta X_h(s)\Delta X_l(s) + \mathbf{c}_{gl}(s)\Delta X_h(s)\Delta X_k(s) \\
& + \mathbf{c}_{hk}(s)\Delta X_g(s)\Delta X_l(s) + \mathbf{c}_{hl}(s)\Delta X_g(s)\Delta X_k(s)] .
\end{aligned}$$

Corollary 2.1 When $p = 1$ in Theorem 2.1, the asymptotic variance V_{gg} is given by

$$V_{gg} = 2 \left[\int_0^1 \mathbf{c}_{gg}^2(s) ds + 2 \sum_{0 < s \leq 1} \mathbf{c}_{gg}(s)(\Delta X_g(s))^2 \right] . \quad (11)$$

The notable point is the fact that the asymptotic distribution and limiting variance-covariances of the SIML estimator have the same forms as the ones of the realized volatility and co-volatilities when there is no noise term if we replace n by m_n , which is dependent on n . It is the key fact to obtain the results of asymptotic properties of the characteristic roots and vectors from the estimated QV in the next section. However, in order to deal with the micro-market noise, we need to use a smaller order $m = [n^\alpha]$ ($0 < \alpha < 0.5$ or 0.4), than n .

3 Use of characteristic roots and vectors

3.1 Reduced rank condition for quadratic covariation

One of important observations on the asset price movements has been the empirical observation that although there are many financial assets traded in markets, many of them move in similar ways with their trends, volatilities and jumps. Then, there is a question how to cope with many asset prices when the number of latent factors of volatilities or quadratic variation of asset prices is less than p , which is the dimension of observed prices. In this section, we consider the case when the underlying continuous time stochastic process is a p -dimensional Itô semimartingale and the number of latent factors of quadratic variation q_x is less than p .

We set the rank condition for the volatilities and co-volatilities:

Assumption 2 Let $r_x = p - q_x (> 0)$ and assume that there exists a $p \times r_x$ ($1 \leq r_x < p$) matrix $\mathbf{B}$ with rank r_x such that

$$\mathbf{B}' \left[\mathbf{X}(t) - \mathbf{X}(0) - \int_0^t \mathbf{b}(s) ds \right] = \mathbf{O} \quad (0 \leq t \leq 1) \quad (12)$$

with probability one. ($\mathbf{B}$ can be random.)

This condition corresponds to the case when the number of latent factors is less than the observed dimension and the latent random components are in the sub-space of $\mathbf{R}^{q_x}$ ($q_x < p$). A related condition is

$$\text{rank}(\Sigma_x) = \text{rank} \left[\int_0^1 \mathbf{c}(s) ds + \sum_{0 \leq s \leq 1} (\Delta \mathbf{X}(s))(\Delta \mathbf{X}(s))' \right] = q_x < p \quad (13)$$

with probability 1. Equation (13) implies that there exists $p \times r_x$ (nonzero) matrix $\mathbf{B}$ consisting of r_x (nonzero) vectors such that

$$\Sigma_x \mathbf{B} = \mathbf{O}, \quad (14)$$

where $\Sigma_x = (\Sigma_{gh}^{(x)})$.

In statistical multivariate analysis and the multivariate errors-in-variables models, the conditions in (12) and (13) have the similar aspect as the reduced rank regression problem or the test of dimensionality. (See Anderson (1984), Anderson (2003) and Fujikoshi et al. (2010), for instance.) The new feature in our formulation is the fact that we are dealing with the continuous-time stochastic process as the latent process, while we have discrete observations with measurement errors.

Since Assumption 2 is restrictive, there can be several ways to relax this condition. For instance, a more general formulation has been considered by Jacod and Podolskij (2013) and Fissler and Podolskij (2017), which are called the test of maximum rank. To avoid a substantial complication, however, we use Assumption 2 in our development. The statistical procedure under this condition becomes considerably simpler than the general case, and it may be enough for applications as we shall illustrate in Sect. 5.

In the present situation, if we take $m_n = [n^\alpha]$ and $0 < \alpha < 0.5$, then from Theorem 2.1 as $n \rightarrow \infty$

$$\hat{\Sigma}_x - \Sigma_m \xrightarrow{P} \mathbf{O}, \quad (15)$$

where $\Sigma_m = (\Sigma_{gh,m})$,

$$\Sigma_m = \Sigma_x + a_m \Sigma_v = (\Sigma_{gh}^{(x)} + a_m \Sigma_{gh}^{(v)}) \quad (16)$$

and

$$a_m = \frac{1}{m_n} \sum_{k=1}^m a_{kn} \quad (17)$$

and $a_{kn} = 4n \sin^2[\frac{\pi}{2}(\frac{2k-1}{2n+1})]$ ($k = 1, \dots, n$).

We also use

$$a_m(2) = \frac{1}{m_n} \sum_{k=1}^m a_{kn}^2 \quad (18)$$

By using the fact that $a_m \rightarrow 0$ as $m_n = n^\alpha \rightarrow \infty$ for $0 < \alpha < 0.5$, and the relation $\sin x \sim x - (1/6)x^3 + o(x^3)$ when x is small, it is straightforward to obtain the next result. (We often use m instead of m_n in the following analysis.)

Lemma 3.1 We set $a_m = (1/m) \sum_{k=1}^m a_{kn}$ and $a_m(2) = (1/m) \sum_{k=1}^m a_{kn}^2$. Then, as $n \rightarrow \infty$ and $m \rightarrow \infty$,

$$\frac{n}{m^2} a_m = \left(\frac{n}{m^2} \right) \frac{1}{m} \sum_{k=1}^m a_{kn} \sim \pi^2 \int_0^1 s^2 ds \quad (19)$$

and

$$\frac{n^2}{m^4} a_m(2) = \frac{1}{m} \sum_{k=1}^m \left(\frac{n^2}{m^4} \right) [a_{kn}]^2 \sim \pi^4 \int_0^1 s^4 ds. \quad (20)$$

In the following derivations, we will investigate the case as if Σ_v is known and $|\Sigma_v| \neq 0$. The results with unknown Σ_v do not depend on this situation if we use a consistent estimator of Σ_v . The variance-covariance matrix Σ_v can be consistently estimated by $\hat{\Sigma}_v$ in (7) when we take $l_n = [n^\beta]$ ($0 < \alpha < \beta < 1$). We can take β being slightly less than 1 and $\hat{\Sigma}_v = \Sigma_v + O_p(\frac{1}{\sqrt{l_n}})$ such that the effects of estimating Σ_v are negligible.

3.2 Asymptotic properties of characteristic vectors

Let the characteristic equation be

$$|\mathbf{G}_m - \lambda \hat{\Sigma}_v| = 0, \quad (21)$$

and $\hat{\mathbf{B}}$, the estimator of $\mathbf{B}$ in (12) and (14), is given by

$$[\mathbf{G}_m \hat{\mathbf{B}} - \hat{\Sigma}_v \hat{\mathbf{B}} \mathbf{\Lambda}] = \mathbf{O}, \quad (22)$$

where $\mathbf{G}_m (= \hat{\Sigma}_x)$ in (5), $\hat{\Sigma}_v$ in (7), λ_i ($i = 1, \dots, p$) are the characteristic roots, $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_{r_x})$ with $0 \leq \lambda_1 \leq \dots \leq \lambda_p$. For the resulting convenience, we take a $p \times r_x$ ($p = r_x + q_x$)

$$\hat{\mathbf{B}} = \begin{bmatrix} \mathbf{I}_{r_x} \\ -\hat{\mathbf{B}}_2 \end{bmatrix}$$

for a normalization of characteristic vectors.

We will use Σ_v instead of $\hat{\Sigma}_v$ and consider $|\mathbf{G}_m - \lambda \Sigma_v| = 0$ in the following derivation by the argument we have discussed.

Instead of Σ_v as a metric, we can take $\mathbf{H}$, which is any positive (known) definite matrix such as the identity matrix $\mathbf{I}_p$. Then, the following derivation and the resulting expression become slightly complicated than the case with Σ_v , and we shall see its consequence at the end of Sect. 3.4.

We take the probability limit of the determinantal equation

$$|\text{plim}_{m \rightarrow \infty} (\mathbf{G}_m - a_m \Sigma_v) - (\text{plim}_{m \rightarrow \infty} \lambda - a_m) \Sigma_v| = 0. \quad (23)$$

The rank of Σ_x is q_x , which is less than p , and $a_m = O(m^2/n)$. Although $\mathbf{G}_m - \Sigma_x = O_p(1/\sqrt{m})$ from (61), we find that the dominant term for the determinantal equation (21) under Assumption 2 should be $\mathbf{G}_m - (\Sigma_x + a_m \Sigma_v) = O_p(\frac{a_m}{\sqrt{m}}) = O_p(\frac{\sqrt{m^3}}{n})$ and we have

$$\lambda_i - a_m \xrightarrow{p} 0 \quad (i = 1, \dots, r_x) \quad (24)$$

if $\frac{\sqrt{m^3}}{n} \rightarrow 0$ as $n \rightarrow +\infty$.

Then,

$$[\text{plim}_{n \rightarrow \infty}(\mathbf{G}_m - a_m \Sigma_v)][\text{plim}_{n \rightarrow \infty} \hat{\mathbf{B}} - \mathbf{B}] = \mathbf{O} . \quad (25)$$

By multiplying $\Pi'_*(q_x \times p)$ from the left hand to

$$\Pi'_*[\text{plim}_{n \rightarrow \infty}(\mathbf{G}_m - a_m \Sigma_v)]\text{plim}_{n \rightarrow \infty} [\hat{\mathbf{B}} - \mathbf{B}] = \mathbf{O} ,$$

such that we can take $\Pi'^* \text{plim}_{n \rightarrow \infty} \mathbf{G}_m [\mathbf{I}_{q_x} \mathbf{O}]$ is non-singular. By using the facts that the rank is q_x and the normalization of $\mathbf{B}$, we find

$$\text{plim}_{n \rightarrow \infty} \hat{\mathbf{B}} = \mathbf{B} . \quad (26)$$

In order to proceed the further step to evaluate the limiting random variables, we use the $\mathbf{K}_n$ -transformation in (6), and we decompose the resulting random variables

$$\mathbf{z}_k = \mathbf{x}_k^* + \mathbf{v}_k^* = \mathbf{x}_k^* + \sqrt{a_{kn}} \mathbf{u}_k^* \quad (k = 1, \dots, m)$$

with $\mathcal{E}(\mathbf{u}_k^* \mathbf{u}_k^{*'}) = \Sigma_v$ and

$$\mathbf{G}_m = \frac{1}{m} \sum_{k=1}^m (\mathbf{x}_k^* + \sqrt{a_{kn}} \mathbf{u}_k^*)(\mathbf{x}_k^* + \sqrt{a_{kn}} \mathbf{u}_k^*)' , \quad (27)$$

where the $p \times 1$ random vectors $\mathbf{x}_k^*$ and $\mathbf{v}_k^*$ are defined by $(\mathbf{x}_k^{*'}) = \mathbf{K}_n(\mathbf{x}'_k)$ and $(\mathbf{v}_k^{*'}) = \mathbf{K}_n(\mathbf{v}'_k)$, which are $n \times p$ matrices, and $\mathbf{u}_k^{*'} = a_{kn}^{-1/2} \mathbf{v}_k^{*'}$.

Under the null-hypothesis $H_0 : \Sigma_x \mathbf{B} = \mathbf{O}$, we have $\mathbf{B}' \Sigma_x \mathbf{B} = \mathbf{O}$ ($r_x \times r_x$ zero matrix). We consider the representation that

$$\mathbf{B}' \mathbf{G}_m = \frac{1}{m} \sum_{k=1}^m \mathbf{B}' (\mathbf{x}_k^* + \sqrt{a_{kn}} \mathbf{u}_k^*)(\mathbf{x}_k^* + \sqrt{a_{kn}} \mathbf{u}_k^*)' .$$

Let $\beta_j (= (\beta_{hj}))$ be the j th column vector of $\mathbf{B}$ ($j = 1, \dots, r_x$) and

$$\sqrt{m}[\mathbf{G}_m - \Sigma_m] \beta_j = \sqrt{m} \left[\sum_{h=1}^p (\hat{\Sigma}_{gh}^{(x)} - \Sigma_{gh,m}) \beta_{hj} \right]_g , \quad (28)$$

where $\Sigma_m = \Sigma_x + a_m \Sigma_v (= (\Sigma_{gh,m}))$ and $a_m = (1/m) \sum_{k=1}^m a_{kn}$.

Since we have the relations $\sum_{h=1}^p \Sigma_{gh}^{(x)} \beta_{hj} = 0$ for $g, j = 1, \dots, p$ under the rank condition, we decompose

$$\begin{aligned}
[\mathbf{G}_m - \boldsymbol{\Sigma}_m] \boldsymbol{\beta}_j &= \frac{1}{m} \sum_{k=1}^m (\mathbf{x}_k^* \mathbf{x}_k^{*'} - \boldsymbol{\Sigma}_x) \boldsymbol{\beta}_j + \frac{1}{m} \sum_{k=1}^m \sqrt{a_{kn}} \mathbf{u}_k^* (\mathbf{x}_k^{*'} \boldsymbol{\beta}_j) \\
&\quad + \frac{1}{m} \sum_{k=1}^m \sqrt{a_{kn}} \mathbf{x}_k^* (\mathbf{u}_k^{*'} \boldsymbol{\beta}_j) + \frac{1}{m} \sum_{k=1}^m a_{kn} (\mathbf{u}_k^* \mathbf{u}_k^{*'} - \boldsymbol{\Sigma}_v) \boldsymbol{\beta}_j.
\end{aligned} \quad (29)$$

Then, we need to evaluate the order of four terms in the above decomposition. Under Assumption 2, we need to evaluate the last two terms. The third term of (29) is $O_p(\sqrt{a_m/m}) (= O_p(\sqrt{m/n}))$, and the order of the last term is $O_p(\sqrt{a_m(2)/m}) (= O_p(\sqrt{m^3/n^2}))$ by using Lemma 3.1. Hence, we find that the dominant term is the third term

$$\mathbf{g}_m = \frac{1}{m} \sum_{k=1}^m \sqrt{a_{kn}} \mathbf{x}_k^* (\mathbf{u}_k^{*'} \boldsymbol{\beta}_j) \quad (30)$$

as $n \rightarrow \infty$. We summarize the result of order evaluation, and a brief derivation is in Appendix.

Lemma 3.2 *We take $m = [n^\alpha]$ with $0 < \alpha < 1$. Under Assumptions 1 and 2, as $m, n \rightarrow \infty$ while $m/n \rightarrow 0$, we have*

$$\sqrt{\frac{m}{a_m}} [\mathbf{g}_m] = O_p(1). \quad (31)$$

Now we shall consider the asymptotic distribution of the characteristic vectors in (22). Let an $r_x \times r_x$ diagonal matrix be $\boldsymbol{\Lambda} = (\text{diag}(\lambda_i))$ and

$$\mathbf{G}_m \hat{\mathbf{B}} - \boldsymbol{\Sigma}_v \hat{\mathbf{B}} [a_m \mathbf{I}_{r_x} + (\boldsymbol{\Lambda} - a_m \mathbf{I}_{r_x})] = \mathbf{O}, \quad (32)$$

which can be written as:

$$[\mathbf{G}_m - a_m \boldsymbol{\Sigma}_v] \hat{\mathbf{B}} = \boldsymbol{\Sigma}_v \hat{\mathbf{B}} [\boldsymbol{\Lambda} - a_m \mathbf{I}_{r_x}]. \quad (33)$$

By multiplying $\hat{\mathbf{B}}'(r_x \times r_x)$ from the left hand to (33), we find

$$\hat{\mathbf{B}}' \boldsymbol{\Sigma}_v \hat{\mathbf{B}} [\boldsymbol{\Lambda} - a_m \mathbf{I}_{r_x}] = \hat{\mathbf{B}}' [\mathbf{G}_m - a_m \boldsymbol{\Sigma}_v] \hat{\mathbf{B}}.$$

Also by multiplying $[\mathbf{O}, \mathbf{I}_{q_x}]$ ($q_x \times p$ matrix) from the left hand to (33), the resulting equation can be rewritten as:

$$[\mathbf{O}, \mathbf{I}_{q_x}] [\mathbf{G}_m - a_m \boldsymbol{\Sigma}_v] [\mathbf{B} + (\hat{\mathbf{B}} - \mathbf{B})] = [\mathbf{O}, \mathbf{I}_{q_x}] \boldsymbol{\Sigma}_v [\mathbf{B} + (\hat{\mathbf{B}} - \mathbf{B})] [\boldsymbol{\Lambda} - a_m \mathbf{I}_{r_x}]. \quad (34)$$

Then, we evaluate the order of each terms of the above equation given that $\hat{\mathbf{B}} - \mathbf{B} \xrightarrow{p} \mathbf{O}$ as $n \rightarrow \infty$. If we have the condition $m/n \rightarrow 0$ as $n \rightarrow \infty$, it is asymptotically equivalent to

$$[\mathbf{O}, \mathbf{I}_{q_x}] \sqrt{\frac{n}{m}} [\mathbf{G}_m - a_m \boldsymbol{\Sigma}_v] \mathbf{B} = [\mathbf{O}, \mathbf{I}_{q_x}] \boldsymbol{\Sigma}_x \begin{bmatrix} \mathbf{O} \\ \mathbf{I}_{q_x} \end{bmatrix} \sqrt{\frac{n}{m}} (\hat{\mathbf{B}}_2 - \mathbf{B}_2) + o_p(1). \quad (35)$$

[The above method has been standard in the statistical multivariate analysis. See Anderson (2003), or Fujikoshi et al. (2010).]

Then, we find the normalization factor c_m as

$$c_m = \sqrt{\frac{m}{a_m}} = O\left(\sqrt{\frac{n}{m}}\right), \quad (36)$$

which goes to infinity as $n \rightarrow \infty$ when $m/n \rightarrow 0$ as $n \rightarrow \infty$.

By noting that (24) is true when $0 < \alpha < 2/3$, we summarize the asymptotic order of $c_m(\hat{\mathbf{B}}_2 - \mathbf{B}_2)$ as the next proposition.

Theorem 3.1 *Suppose Assumptions 1 and 2 are satisfied. We take $m = m_n = [n^\alpha]$, $l_n = [n^\beta]$ with $0 < \alpha < \beta < 1$, $0 < \alpha < 2/3$, and $\hat{\boldsymbol{\Sigma}}_v$. Let $c_m \rightarrow \infty$ as $n \rightarrow \infty$. Then, as $n \rightarrow \infty$, we have that*

$$c_m [\hat{\mathbf{B}}_2 - \mathbf{B}_2] = O_p(1). \quad (37)$$

3.3 Asymptotic properties of characteristic roots

Next, we investigate the limiting distribution of the characteristic roots of $\mathbf{\Lambda}$ and the related statistics.

From (29), let

$$\begin{aligned} \sqrt{m} \mathbf{B}' (\mathbf{G}_m - \boldsymbol{\Sigma}_m) \mathbf{B} &= \sqrt{m} \left[\frac{1}{m} \sum_{k=1}^m \mathbf{B}' (\mathbf{x}_k^* \mathbf{x}_k^{*'} - \boldsymbol{\Sigma}_x) \mathbf{B} + \frac{1}{\sqrt{m}} \sum_{k=1}^m \sqrt{a_{kn}} \mathbf{B}' \mathbf{u}_k^* (\mathbf{x}_k^{*'} \mathbf{B}) \right. \\ &\quad \left. + \frac{1}{\sqrt{m}} \sum_{k=1}^m \sqrt{a_{kn}} \mathbf{B}' \mathbf{x}_k^* \mathbf{u}_k^{*'} \mathbf{B} + \frac{1}{\sqrt{m}} \sum_{k=1}^m a_{kn} (\mathbf{u}_k^* \mathbf{u}_k^{*'} - \boldsymbol{\Omega}_v) \right] \end{aligned}$$

and $\boldsymbol{\Omega}_v = \mathbf{B}' \boldsymbol{\Sigma}_v \mathbf{B}$ ($= (\omega_{gh})$).

Under Assumption 2, we only need to evaluate the last term, which can be written as

$$\mathbf{E}_m = \frac{1}{\sqrt{m}} \sum_{k=1}^m a_{kn} (\mathbf{u}_k^* \mathbf{u}_k^{*'} - \boldsymbol{\Omega}_v). \quad (38)$$

By applying CLT to (38), under the assumption of existence of 4th order moments of noise terms, we have the asymptotic normality, which is summarized in the next lemma and the derivation of a CLT with asymptotic covariances is given in Appendix.

Lemma 3.3 Let $\mathbf{\Omega}_v = \mathbf{B}'\mathbf{\Sigma}_v\mathbf{B} = (\omega_{gh})$. Under Assumptions 1 and 2, the asymptotic distribution of each elements of $\sqrt{\frac{1}{a_m(2)}}\mathbf{E}_m = (e_{gh}^*)$ is asymptotically normal as $m, n \rightarrow \infty$ while $m/n \rightarrow 0$.

The asymptotic covariance of e_{gh}^* and e_{kl}^* ($g, h, k, l = 1, \dots, r_x$) is given by $\omega_{gk}\omega_{hl} + \omega_{gl}\omega_{hk}$.

Because of Lemmas 3.1 and A.3 in Appendix,

$$\mathbf{B}'(\mathbf{G}_m - \mathbf{\Sigma}_m)\mathbf{B} = O_p\left(\sqrt{\frac{a_m(2)}{m}}\right) = O_p\left(\frac{a_m}{\sqrt{m}}\right) \quad (39)$$

and

$$\mathbf{B}'\mathbf{G}_m\begin{pmatrix} \mathbf{O} \\ \mathbf{I}_q \end{pmatrix}(\hat{\mathbf{B}}_2 - \mathbf{B}_2) = O_p\left(\sqrt{\frac{a_m}{m}}\right) \times O_p\left(\sqrt{\frac{a_m}{m}}\right) = O_p\left(\frac{a_m}{m}\right);$$

By using the decomposition $\hat{\mathbf{B}} = \mathbf{B} + [\hat{\mathbf{B}} - \mathbf{B}]$, we can evaluate as

$$\begin{aligned} & \hat{\mathbf{B}}'[\mathbf{G}_m - a_m\mathbf{\Sigma}_v]\hat{\mathbf{B}} \\ &= \mathbf{B}'[\mathbf{G}_m - a_m\mathbf{\Sigma}_v]\mathbf{B} - \mathbf{B}'[\mathbf{G}_m - a_m\mathbf{\Sigma}_v]\begin{bmatrix} \mathbf{O} \\ \mathbf{I}_{q_x} \end{bmatrix}\mathbf{\Sigma}_*^{-1}[\mathbf{O}, \mathbf{I}_{q_x}][\mathbf{G}_m - a_m\mathbf{\Sigma}_v]\mathbf{B} \\ &= O_p\left(\frac{a_m}{\sqrt{m}}\right) + O_p\left(\frac{m}{n}\right), \end{aligned} \quad (40)$$

where

$$\mathbf{\Sigma}_* = [\mathbf{O}, \mathbf{I}_{q_x}]\mathbf{\Sigma}_x\begin{bmatrix} \mathbf{O} \\ \mathbf{I}_{q_x} \end{bmatrix}.$$

Since the first term of (40) is $O_p(\frac{a_m}{\sqrt{m}}) = O_p(\frac{m\sqrt{m}}{n})$ and the second term is $O_p(\frac{a_m}{m}) = O_p(\frac{m}{n})$, the second term of the right-hand side is asymptotically negligible. Hence, the limiting distribution of

$$\sqrt{m}(\mathbf{B}'\mathbf{\Sigma}_v\mathbf{B})\left[\left(\frac{1}{a_m}\right)\mathbf{\Lambda} - \mathbf{I}_{r_x}\right] \sim \frac{\sqrt{m}}{a_m}\mathbf{B}'(\mathbf{G}_m - \mathbf{\Sigma}_m)\mathbf{B} \quad (41)$$

is asymptotically normal by applying CLT to (38).

We write $\mathbf{E}_m = (e_{ij}) = \frac{1}{\sqrt{m}} \sum_{k=1}^m a_{kn}(\mathbf{u}_k^* \mathbf{u}_k^{*'} - \mathbf{\Omega}_v)$ and $\mathbf{D} = \mathbf{\Omega}_v^{-1}\mathbf{E} (= (d_{ij}))$. Then, by using Lemmas 3.3 and A.3 in Appendix,

$$\begin{aligned}
d_{ij} &= \text{Cov} \left[\sum_{k=1}^r \omega^{ik} e_{ki}, \sum_{k'=1}^r \omega^{jk'} e_{k'j} \middle| \mathcal{F}^{(0)} \right] \\
&= \frac{a_m(2)}{a_m^2} \sum_{k,k'=1}^r \omega^{ik} \omega^{jk'} [\omega_{kk'} \omega_{ij} + \omega_{kj} \omega_{k'i}] + o_p(1) \\
&= \frac{a_m(2)}{a_m^2} [\delta(i,j) + \omega_{ij} \omega^{ij}] + o_p(1),
\end{aligned} \tag{42}$$

for $e_{kj} = (\sqrt{m}/a_m)(\mathbf{B}'(\mathbf{G}_m - \mathbf{\Sigma}_m)\mathbf{B})_{kj}$ and $\mathbf{\Omega}_v^{-1} = (\omega^{ij})$. (We shall ignore the last term $o_p(1)$ of (42) in the following expression for the resulting simplicity.)

When we use a consistent estimator of $\mathbf{\Sigma}_v$, the resulting expression of the limiting distribution becomes simple, which may be useful in practice. By using the fact that $\sqrt{a_m(2)}/a_m \sim 3/\sqrt{5}$ and Lemma 3.3, we obtain the next result on the asymptotic distributions of the smaller characteristic roots λ_i ($i = 1, \dots, r_x$).

Theorem 3.2 *Assume the conditions on the Itô semimartingale in (1) and (2) as Theorem 3.1. We take $m = m_n = [n^\alpha]$ ($0 < \alpha < 2/3$), $l_n = [n^\beta]$ with $0 < \alpha < \beta < 1$ and $\hat{\mathbf{\Sigma}}_v$ for $\mathbf{\Sigma}_v$. As $n \rightarrow \infty, m \rightarrow \infty$,*

$$\frac{\sqrt{m}}{a_m} [\lambda_i - a_m] \xrightarrow{d} N\left(0, \frac{9}{5} d_{ii}\right) \tag{43}$$

for $i = 1, \dots, r_x$ ($= p - q_x$) and $[\sqrt{m}/a_m][\lambda_i - a_m]$ are asymptotically normal jointly. The covariances of $[\sqrt{m}/a_m][\lambda_i - a_m]$ and $[\sqrt{m}/a_m][\lambda_j - a_m]$ ($i, j = 1, \dots, r_x$) are given by

$$d_{ij} = \frac{9}{5} [\delta(i,j) + \omega_{ij} \omega^{ij}].$$

Let

$$\lambda_i^* = \frac{\sqrt{m}}{a_m} [\lambda_i - a_m] \quad (i = 1, \dots, r_x). \tag{44}$$

Since the effect of estimating $\mathbf{\Sigma}_v$ is asymptotically negligible, we find the asymptotic variance as

$$\begin{aligned}
\text{AVar} \left[\sum_{i=1}^{r_x} \lambda_i^* \right] &= \mathcal{E} \left[\sum_{i,i'=1}^{r_x} \lambda_i^* \lambda_{i'}^* \right] \\
&= \left[\frac{a_m(2)}{a_m^2} \sum_{i,i'=1}^{r_x} [\delta(i,i') + \omega_{ii'} \omega^{ii'}] \right] = \frac{9}{5} 2r_x.
\end{aligned} \tag{45}$$

Then, we summarize our main result on the trace-statistic $\sum_{i=1}^{r_x} \lambda_i$, which will be used as the key statistic for the application.

Theorem 3.3 Assume the conditions in Theorem 3.2. We take $m = m_n = [n^\alpha]$ ($0 < \alpha < 2/3$), $l_n = [n^\beta]$ with $0 < \alpha < \beta < 1$ and $\hat{\Sigma}_v$. As $n \rightarrow \infty$ and $m \rightarrow \infty$,

$$\sqrt{\frac{m}{a_m(2)}} \left[\sum_{i=1}^{r_x} (\lambda_i - a_m) \right] \xrightarrow{d} N(0, 2r_x). \quad (46)$$

and

$$\frac{m}{a_m(2)} \frac{1}{2r_x} \left[\sum_{i=1}^{r_x} (\lambda_i - a_m) \right]^2 \xrightarrow{d} \chi^2(1). \quad (47)$$

When the true rank r_x^* in (12) is less than r_x , which is the number of roots for constructing the statistic in Theorem 3.3, the probability limit of the r_x -th sample characteristic root is nonzero. Hence, we immediately obtain the following result.

Corollary 3.1 Assume the conditions in Theorem 3.2 except the condition that the true rank $q_x^* > q_x = p - r_x$ in (12). Then,

$$\frac{m}{a_m(2)} \frac{1}{2r_x} \left[\sum_{i=1}^{r_x} (\lambda_i - a_m) \right]^2 \xrightarrow{p} \infty. \quad (48)$$

3.4 On eigenvalues in the metric $\mathbf{H}$

If we take a non-singular (known) matrix $\mathbf{H}$ in (21) and (22) instead of $\hat{\Sigma}_v$, then we need a restrictive condition and the derivation in Sect. 3.3 should be modified slightly. It is because we cannot use (23), but the resulting derivation is similar to the one in Sect. 3.3. When we use any $\mathbf{H}$, let

$$\Lambda_0^{**} = (\mathbf{B}' \mathbf{H} \mathbf{B})^{-1} \mathbf{Q}_v (= (\lambda_{ij}^{**})), \quad (49)$$

which corresponds to the probability limits of $(1/a_m)\mathbf{A}$.

Then, we can replace $\mathbf{B}$ and $\mathbf{Q}_v (= \mathbf{B}' \hat{\Sigma}_v \mathbf{B})$ by their consistent estimators. In this case, however, the requirement on m is different from Theorems 3.2 and 3.3. We present the following result, but we omit the details of derivation because the basic argument is similar to the one in Sect. 3.3.

Proposition 3.1 Assume the conditions in Theorem 3.2. We take $m = m_n = [n^\alpha]$ ($0 < \alpha < 1/2$) and a non-singular (constant) matrix $\mathbf{H}$ instead of $\hat{\Sigma}_v$. As $n \rightarrow \infty$, $m \rightarrow \infty$ and

$$\frac{m}{a_m(2)} \frac{1}{r_*} \left[\sum_{i=1}^{r_x} (\lambda_i - a_m \lambda_{ii}^{**}) \right]^2 \xrightarrow{d} \chi^2(1), \quad (50)$$

where $r_* = \sum_{i,i'=1}^{r_x} \text{Cov}(U_{ii}^*, U_{i'i'}^*)$ and the covariances are given as Lemma A.3 in Appendix with U_{ii}^* .

It may be convenient to take $\mathbf{H} = \mathbf{I}_p$. But there is a complication in the expression of asymptotic distribution as well as we need an additional estimation of unknown parameters from observations.

3.5 A procedure of detecting the number of factors of quadratic variation

It is straightforward to develop the testing procedure for the hypothesis $H_0 : r_x = r_0$ ($p > r_0 \geq 1$ is a specified number) against $H_A : r_x = r_0 + 1$. It may be reasonable to use the r_0 th smaller characteristic root, and the rejection region can be constructed by the limiting normal or χ^2 distribution under H_0 . (H_0 corresponds to the case of $q_x = p - r_0$ while H_A corresponds to $q_x = p - (r_0 + 1)$.) Hence, it may be natural to use the sum of smaller characteristic roots as

$$R_0 = \sum_{i=1}^{r_0} (\lambda_i - a_m) \quad (51)$$

where $0 \leq \lambda_1 \leq \dots \leq \lambda_p$. From Corollary 3.1, we can use

$$T_n(r_0) = \frac{m}{a_m(2)} \frac{1}{2r_0} \left[\sum_{i=1}^{r_0} (\lambda_i - a_m) \right]^2 \quad (52)$$

as the test statistics for estimating the number of latent factors of the underlying Itô semimartingale. The rejection region with $1 - \alpha_n$ ($\alpha_n = \alpha / \log m_n$ for some constant $\alpha \in (0, 1)$) significance level should be

$$T_n(r_0) \geq \chi_{1-\alpha_n}^2(1), \quad (53)$$

where $\chi_{1-\alpha_n}^2(1)$ is the $(1 - \alpha_n)$ -quantile of $\chi^2(1)$. Under H_A , the $r_0 + 1$ th characteristic root $\lambda_{r_0+1} \rightarrow \infty$ and the test should be consistent, that is, we can consistently estimate the true number of latent factors $q_x^* = p - r_x^*$.

Formally, we employ the following stopping rule for the proposed sequential test.

1. Compute the test statistics $T_n(p)$. If $T_n(p) < \chi_{1-\alpha_n}^2(1)$, we conclude $r_x = p$. (i.e., $\hat{q}_x^* = 0$).
If $T_n(p) \geq \chi_{1-\alpha_n}^2(1)$, we proceed to the next step to test $H_0 : r_x = p - 1$ against $H_A : r_x = p$ by using $T_n(p - 1)$.
2. We iterate the test $H_0 : r_x = r_0 - 1$ against $H_A : r_x = r_0$ ($r_0 = p, p - 1, \dots, 2$) sequentially until the null hypothesis is accepted.
3. We finish the sequential test and conclude $r_x = r_0 - 1$ (i.e., $\hat{q}_x^* = p - (r_0 - 1)$) at the time when $H_0 : r_x = r_0 - 1$ is accepted for the first time.
4. When the null hypothesis $H_0 : r_x = 1$ is rejected, we conclude $r_x = 0$ (i.e., $\hat{q}_x^* = p$).

We can show the validity of the proposed method, which is summarized in the next proposition and the proof is given in Appendix.

Theorem 3.4 *Under the same conditions in Theorem 3.2, we have that $\hat{q}_x^* \xrightarrow{P} q_x^*$ as $n \rightarrow \infty$.*

Our method corresponds to an extension of the standard statistical method to the case when we have latent continuous stochastic process and to detect the number of factors in the statistical multivariate analysis. See Anderson (1984), Anderson (2003), Robin and Smith (2000), and Fujikoshi et al. (2010).

In the framework of traditional hypothesis testing, there is a multiple testing problem when we use a sequence of test statistics with a given significance level. In the time series analysis, there is a literature on model selection for determining the number of factors by Akaike's information criterion (AIC). See Kitagawa (2010), for instance. These problems are important for practical purposes, which will be explored in another occasion.

4 Simulations

In this section, we give some simulation results on the characteristic roots and test statistics when $p = 10$ (the observed dimension). We also discuss the finite sample properties of the characteristic roots and test statistics we have developed in the previous sections.

4.1 Simulated models

First, we simulate the latent process with three-dimensional Itô semimartingales. Let $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}(t))_{t \geq 0} = (\tilde{X}_1(t), \tilde{X}_2(t), \tilde{X}_3(t))'_{t \geq 0}$ be the vector of Itô semimartingale satisfying

$$d\tilde{X}_g(t) = \sigma_g(t) dW_g(t), \quad g = 1, 2, \quad (54)$$

$$d\tilde{X}_3(t) = Z(t) dN(t), \quad (55)$$

where $\mathbf{W} = (W_1, W_2)'$ is the two-dimensional (standard) Brownian motion vector, N is the Poisson process with intensity 10. We assume that N is independent of $\mathbf{W}$ and $Z = (Z(t))_{t \geq 0}$ is the jump sizes with $Z(t) \sim N(0, 5^{-2})$. (See Cont and Tankov (2004) for the generation of jump processes.)

For the volatility process σ of the diffusion part, we set

$$d(\sigma_g(t))^2 = a_g(\mu_g - (\sigma_g(t))^2)dt + \kappa_g \sigma_g(t) dW^\sigma(t), \quad g = 1, 2, \quad (56)$$

where $\sigma_1(t)$ and $\sigma_2(t)$ are independent, $\mathbf{W}^\sigma = (W^{\sigma_1}, W^{\sigma_2})'$ is the two dimensional (standard) Brownian motion. We have set $a_1 = 2$, $a_2 = 3$, $\mu_1 = 0.8$, $\mu_2 = 0.7$,

$\kappa_1 = \kappa_2 = 0.5$, $\mathcal{E}[dW_j(t)dW_j(t)] = \rho_j dt$, and $\rho_1 = \rho_2 = -0.5$. In our simulation, we consider the following two models:

$$\text{Model 1 : } \mathbf{Y}(t) = \mathbf{\Gamma}_1(\tilde{X}_1(t), \tilde{X}_2(t))' + \mathbf{v}(t) \quad (57)$$

and

$$\text{Model 2 : } \mathbf{Y}(t) = \mathbf{\Gamma}_1(\tilde{X}_1(t), \tilde{X}_2(t))' + \mathbf{\Gamma}_2\tilde{X}_3(t) + \mathbf{v}(t). \quad (58)$$

Here, we denote the coefficient matrices ($p \times 2$ and $p \times 1$, respectively) as $\mathbf{\Gamma}_1 = (\gamma_1^{(1)}, \dots, \gamma_1^{(p)})'$ ($\gamma_1^{(j)} = (\gamma_{1,1}^{(j)}, \gamma_{1,2}^{(j)})$) and $\mathbf{\Gamma}_2 = (\gamma_2^{(1)}, \dots, \gamma_2^{(p)})'$, which are sampled as $\gamma_{1,1}^{(j)} \sim U([0.25, 1.75])$, $\gamma_{1,2}^{(j)} \sim U([0.1, 0.25])$ and $\gamma_2^{(j)} \sim U([0.25, 1.75])$, $j = 1, \dots, p$. The observation vectors are

$$\mathbf{Y}(t_i^n) = (Y_1(t_i^n), \dots, Y_p(t_i^n))', \quad i = 1, \dots, n$$

and we set $t_i^n = \frac{i}{n}$ ($i = 1, \dots, n$) and $\Delta = \Delta_n = 1/n$. As the market microstructure noise vectors, we set $\mathbf{v}(t_i^n) = (v_1(t_i^n), \dots, v_p(t_i^n))'$, and use independent Gaussian noises for each component, that is,

$$(v_1(t_i^n), \dots, v_p(t_i^n))' \sim i.i.d.N_p(\mathbf{0}, c_v \mathbf{I}_p) \quad (i = 1, \dots, n),$$

with a pre-specified value c_v . In all simulations, we set $p = 10$ and $n = 20,000$. (Models 1 and 2 can be seen as special cases, which investigated in Li et al. (2017a, b).)

In our simulation, we have set $m_n = 2 \times [n^{0.646}]$ because of the conditions in Theorems 3.2 and 3.3. We have done a large number of simulations, but we report only some results, which seem to be representative cases.

4.2 Simulation results

Let N be the number of Monte Carlo iterations. We plotted the mean value of the eigenvalues of the SIML estimator for the quadratic variation in Figs. 1 and 2. To compute the SIML estimators, we set $m_n = 2 \times [n^{0.646}]$ ($= 2 \times 600$) and $l_n = 1.5m_n$.

$$\hat{\Sigma}_x = \frac{1}{m_n} \sum_{j=1}^{m_n} \mathbf{z}_j \mathbf{z}_j', \quad \hat{\Sigma}_v = \frac{1}{l_n} \sum_{j=n-l_n+1}^n a_{jn}^{-1} \mathbf{z}_j \mathbf{z}_j', \quad (59)$$

where $a_{kn} = 4n \sin^2 \left[\frac{\pi}{2} \left(\frac{2k-1}{2n+1} \right) \right]$.

In the following figures, we set $0 \leq \lambda_1 \leq \dots \leq \lambda_p$ are eigenvalues of $\hat{\Sigma}_v^{-1} \hat{\Sigma}_x$.

In Model-1 and Model-2, we have 10 dimensions observation vectors ($p = 10$). Model-1 has two factor of diffusion type ($q_x = 2, r_x = 8$), while Model-2 has two diffusion type factors and one jump factor ($q_x = 3, r_x = 7$). Figures 1 and 2 show that the estimated characteristic roots reflect the true rank of hidden stochastic process. Figures 3 and 4 shows the distributions of the test statistic we have developed in this paper. In the first case (Fig. 3), there is a clear indication that there are two nonzero characteristic roots, which are far from zero by any meaningful criterion, while in

Fig. 1 Mean of estimated log characteristic roots (log eigenvalues) of Model 1 when $c_p = 10^{-6}$ (left) and $c_p = 10^{-8}$ (right). We set $\Delta = 1/20,000$, $m_n = 2 \times [n^{0.6461}] (= 2 \times 600)$ and $l_n = 1.5m_n$. The number of Monte Carlo iteration is 300

Fig. 2 Mean of estimated log characteristic roots (log eigenvalues) of Model 2 when $c = 10^{-6}$ (left) and $c = 10^{-8}$ (right). We set $\Delta = 1/20,000$, $m_n = 2 \times \lceil n^{0.646} \rceil$ ($= 2 \times 600$) and $l_n = 1.5m_n$. The number of Monte Carlo iteration is 300

Fig. 3 Empirical distributions of test statistic $T_n(r_0)$ of Model 1 when $r_0 = 7$ (left), $r_0 = 8$ (center) and $r_0 = 9$ (right) when $r_x = 8$. We set $\Delta = 1/20,000$, $c_p = 10^{-8}$, $m_n = 2 \times [n^{0.646}]$ ($= 2 \times 600$) and $L_n = 1.5m_n$. The number of Monte Carlo iteration is 300. The red line is the density of the Chi-square distribution with 1 degree of freedom

Fig. 4 Empirical distributions of test statistic $T_n(r_0)$ of Model 2 when $r_0 = 6$ (left), $r_0 = 7$ (center) and $r_0 = 8$ (right) when $r_x = 7$. We set $\Delta = 1/20,000$, $c_p = 10^{-8}$, $m_n = 2 \times \lceil n^{0.646} \rceil$ ($= 2 \times 600$) and $l_n = 1.5m_n$. The number of Monte Carlo iteration is 300. The red line is the density of the Chi-square distribution with 1 degree of freedom

Fig. 5 A qq-plot of eigenvalues (Model-2 when $r_x = 7$)

the second case (Fig. 4), there is a clear indication that there are three nonzero characteristic roots, which are far from zero with any meaningful criterion.

To illustrate the asymptotic normality of characteristic roots obtained by Theorem 3.2, we give one qq-plot as Fig. 5 for Model-2 when ($q_x = 3, r_x = 7$). We have confirmed the result of Theorem 3.2 by using our simulation.

As an illustration of the power of test statistic developed, we report the relative number of rejection of the hypothesis for 10% and 5% levels by using χ^2 -statistic. (The number of replication was 300.) In Model 1 (the null is $r_x = 8$), the size was 0.840 (90%) and it was 0.913 (95%). In Model 2 (the null is $r_x = 7$), the size was 0.853 (90%) and it was 0.917 (95%).

It seems that our method of evaluating the rank condition of latent volatility factors based on the characteristic roots and the SIML estimation detects the number of latent factors properly by our simulations.

5 An empirical example

In this section, we report one empirical data analysis by using the proposed method developed in the previous sections. It is no more than an illustration on our proposed method in this paper. We have used the intra-day observations of top five financial stocks (Mitsubishi UFJ Financial Group, Inc., Mizuho Financial Group, Inc., Nomura Holdings, Inc., Resona Holdings, Inc., and Sumitomo Mitsui Financial Group, Inc.) traded in the Tokyo Stock Exchange (TSE) on January 25 in

Fig. 6 Estimated eigenvalues. In this case, $\Delta = 1/18,000$, $m_n = 2 \times [n^{0.51}] (= 294)$ and $l_n = 1.5m_n$

2016, which may be regarded as a typical 1 day.¹ We have picked 5 major financial stocks listed at TSE because they are actively traded in each day with high liquidity. Hence, we do not have serious disturbing effects due to actual non-synchronous trading process in TSE market. We sub-sampled returns of each asset every 1 s ($\Delta = n^{-1} = 1/18,000$), and we have taken the nearest trading (past) prices at every unit of time.

For the SIML estimation, we have set $m_n = 2 \times [n^{0.51}] (= 294)$ and $l_n = 1.5m_n$. Since all companies belong to the same market division (first section) of TSE, it would be reasonable to expect that the number of factor of these assets is smaller than 5 (i.e., $q_x < 5$). Figure 6 shows the estimated eigenvalues of the quadratic variation of these stocks by using (21) and (22). In this example, we have two large eigenvalues, while there are three smaller eigenvalues and two roots are dominant. Then, we have the statistics $T_n(5) = 91.37832$, $T_n(4) = 40.41634$, $T_n(3) = 5.479696$ and $T_n(2) = 10.60642$. Therefore, at a significance level of $\alpha_n = 0.05/\log m_n$ ($\chi^2_{1-\alpha_n}(1) \approx 6.8635$), the null hypotheses $H_0 : r_x = 5$ and $H_0 : r_x = 4$ are rejected, but the null hypothesis $H_0 : r_x = 3$ is not rejected. In this example, there is a large root and the second larger root is much smaller than the largest root, but we cannot ignore it because other roots are much smaller than two roots. It implies that the quadratic variation has two factors in 1 day.

Figure 7 shows the intra-day movements of five stock prices in the TSE afternoon session of January 25, 2016. (There is a lunch break in Tokyo stock market.) We set the same values for the starting prices because we want to focus on the volatility

¹ We have analyzed stock prices in several typical days and report the result for 1 day. We expect similar results for other days, but the full analysis is beyond scope of the present paper.

Fig. 7 Intra-day movements of 5 stock prices at Tokyo (January 25, 2016)

structure (or quadratic variation) of five asset prices. There is a strong evidence on two types of intra-day movements of stock prices, which is consistent with our data analysis reported.

6 Conclusions

In financial markets, we usually have many assets traded and then it is important to find latent (small number of) factors behind the observed random fluctuations of underlying assets. It is straightforward to detect the number of latent factors by using the SIML method when the true latent stochastic process is the class of Itô semimartingales, and there can be market microstructure noises. Our procedure is essentially the same as the standard method in statistical multivariate analysis except the fact that we have Itô semimartingales as the latent state variables. We have derived the asymptotic properties of characteristic roots and vectors, which are new. Then, it is possible to develop the test statistics for the rank condition. From our limited simulations and an empirical application, our approach works well in practical situations.

There can be possible extensions of our approach we have developed in this paper. Since Assumptions 1 and 2 are restrictive, it would be interesting to explore the mathematical details and find less restrictive conditions when we can lead to useful results. For instance, there are some discussion on the presence of common jumps such as

Jacod and Todorov (2009), Bibinger and Winkelmann (2015), Aït-Sahalia and Jacod (2014) and Kurisu (2018) on this point. Also some comparison with the existing literature would be useful although it requires a further analysis. Also there are several unsolved problems remained. Although we need to choose m_n and l_n in the SIML testing in practice, there is a different aspect from the problem of choosing m_n and l_n in the SIML estimation. The testing power of test procedure will be another unsolved problem although the trace statistic we used may be a natural choice for test procedure.

More importantly, the notable feature of our approach under Assumption 2 is that the method is simple and is useful. We have a promising experiment even in the case when the dimension of observations is not small, say 100. The number of asset prices in actual financial markets is large in practical financial risk management, and there would be a number of empirical applications. These problems are currently under investigation.

Appendix: On mathematical derivations

In Appendix, we give some mathematical details used in the previous sections. For the sake of expository purpose, we refer Kunitomo et al. (2018) as KSK (2018) and repeat some parts of their Chapter 5.

(i) An Outline of the Derivation of Theorem 2.1

(Step 1): We give an outline of derivation, which may be sometimes rather intuitive for the result. The basic method of proof is essentially the same to the one based on the notations, derivations and their extensions given in Chapter 5 of KSK (2018). We first consider the case when $\mathbf{X}$ has no jumps and the volatility functions are deterministic. Then, we will discuss some extra arguments when the volatility functions are stochastic. We will discuss the case when there can be jumps in Step 2 and mention to important points for the stable convergence of limiting random variables as $n \rightarrow \infty$ in Step 3.

Let $\mathbf{x}_k^{*'} = (x_{kj}^*)$ and $\mathbf{v}_k^{*'} = (v_{kj}^*)$ ($k = 1, \dots, n$) be the k th row vector elements of $n \times p$ matrices

$$\mathbf{X}_n^* = \mathbf{K}_n(\mathbf{X}_n - \bar{\mathbf{Y}}_0), \quad \mathbf{v}_n^* = \mathbf{K}_n \mathbf{v}_n, \quad (60)$$

respectively, where we denote $\mathbf{X}_n = (\mathbf{x}_k') = (x_{kg})$, $\mathbf{v}_n = (\mathbf{v}_k') = (v_{kg})$, $\mathbf{Z}_n = (\mathbf{z}_k') = (z_{kg})$ are $n \times p$ matrices with the notations $z_{kg} = x_{kg}^* + v_{kg}^*$ in Section 3. We write z_{kg} as the g th component of $\mathbf{z}_k$ ($k = 1, \dots, n; g = 1, \dots, p$). We use the decomposition of $\mathbf{z}_k' = (z_{kg}^{(x)})$ for investigating the asymptotic distribution of $\sqrt{m_n}[\hat{\Sigma}_x - \Sigma_x] = (\sqrt{m_n}(\hat{\Sigma}_{gh}^{(x)} - \Sigma_{gh}^{(x)}))$ for $g, h = 1, \dots, p$. We decompose

$$\begin{aligned} & \sqrt{m_n}[\hat{\Sigma}_x - \Sigma_x] \\ &= \sqrt{m_n} \left[\frac{1}{m_n} \sum_{k=1}^{m_n} \mathbf{x}_k^* \mathbf{x}_k^{*'} - \Sigma_x \right] + \frac{1}{\sqrt{m_n}} \sum_{k=1}^{m_n} \mathcal{E}[\mathbf{v}_k^* \mathbf{v}_k^{*'}] \\ & \quad + \frac{1}{\sqrt{m_n}} \sum_{k=1}^{m_n} [\mathbf{v}_k^* \mathbf{z}_k^{*'} - \mathcal{E}[\mathbf{v}_k^* \mathbf{v}_k^{*'}]] + \frac{1}{\sqrt{m_n}} \sum_{k=1}^{m_n} [\mathbf{x}_k^* \mathbf{v}_k^{*'} + \mathbf{v}_k^* \mathbf{x}_k^{*'}]. \end{aligned} \quad (61)$$

Then, we can investigate the conditions that three terms except the first one are $o_p(1)$. When these conditions are satisfied, we could estimate the variance and covariance of the underlying processes consistently as if there were no noise terms because other terms can be ignored asymptotically as $n \rightarrow \infty$.

Let $\mathbf{b}_k = (b_{kj}) = h_n^{-1/2} \mathbf{e}'_k \mathbf{P}_n \mathbf{C}_n^{-1} = (b_{kj})$ and $\mathbf{e}'_k = (0, \dots, 1, 0, \dots)$ be an $n \times 1$ vector. We write $v_{kg}^* = \sum_{j=1}^n b_{kj} v_{jg}$ for the noise part and use the relation

$$h_n^{-1} (\mathbf{P}_n \mathbf{C}_n^{-1} \mathbf{C}_n^{-1} \mathbf{P}'_n)_{k,k'} = \delta(k, k') 4n \sin^2 \left[\frac{\pi}{2n+1} \left(k - \frac{1}{2} \right) \right].$$

Because we have $\sum_{j=1}^n b_{kj} b_{k'j} = \delta(k, k') a_{kn}$ and Σ_v is bounded, it is straightforward to find K_1 (a positive constant) such that $\mathcal{E}[(v_{kg}^*)^2] = \mathcal{E}[\sum_{i=1}^n b_{ki} v_{ig} \sum_{j=1}^n b_{kj} v_{jg}] \leq K_1 \times a_{kn}$. By using Lemma 3.1, the second term becomes

$$\frac{1}{\sqrt{m_n}} \sum_{k=1}^{m_n} \mathcal{E}[v_{kn}^*]^2 \leq K_1 \frac{1}{\sqrt{m_n}} \sum_{k=1}^{m_n} a_{kn} = O\left(\frac{m_n^{5/2}}{n}\right),$$

which is $o(1)$ if we set α such that $0 < \alpha < 0.4$. For the fourth term of (61),

$$\mathcal{E} \left[\frac{1}{\sqrt{m_n}} \sum_{k=1}^{m_n} x_{kg}^* v_{kg}^* \right]^2 = \frac{1}{m_n} \sum_{k,k'=1}^{m_n} \mathcal{E} [x_{kg}^* x_{k'g}^* v_{kg}^* v_{k'g}^*] = O\left(\frac{m_n^2}{n}\right).$$

We set $s_{jk} = \cos[\frac{2\pi}{2n+1}(j - \frac{1}{2})(k - \frac{1}{2})]$ for $j, k = 1, 2, \dots, n$ and then, we have the relation $|\sum_{j=1}^n s_{jk} s_{j,k'}| \leq [\sum_{j=1}^n s_{jk}^2] = \frac{n}{2} + \frac{1}{4}$ for any $k \geq 1$. For the third term of (61), we need to consider the variance of

$$(v_{kg}^*)^2 - \mathcal{E}[(v_{kg}^*)^2] = \sum_{j,j'=1}^n b_{kj} b_{kj'} [v_{jg} v_{j'g} - \mathcal{E}(v_{jg} v_{j'g})].$$

Then, by using the assumption on the existence of the fourth-order moments of market microstructure noise terms, we can find a positive constant K_2 such that

$$\begin{aligned} & \mathcal{E} \left[\frac{1}{\sqrt{m_n}} \sum_{k=1}^{m_n} ((v_{kg}^*)^2 - \mathcal{E}[(v_{kg}^*)^2]) \right]^2 \\ &= \frac{1}{m_n} \sum_{k_1, k_2=1}^{m_n} \mathcal{E} \left[\sum_{j_1, j_2, j_3, j_4=1}^n b_{k_1 j_1} b_{k_2 j_2} (v_{j_1 g} v_{j_2 g} - \mathcal{E}(v_{j_1 g} v_{j_2 g})) \right. \\ & \quad \left. \times b_{k_3 j_3} b_{k_4 j_4} (v_{j_3 g} v_{j_4 g} - \mathcal{E}(v_{j_3 g} v_{j_4 g})) \right] \\ & \leq K_2 \frac{1}{m_n} \left[\sum_{k=1}^{m_n} a_{kn} \right]^2 = O\left(\frac{1}{m_n} \times \left(\frac{m_n^3}{n}\right)^2\right), \end{aligned}$$

which is $O(m_n^5/n^2)$. Thus, the third term of (61) is negligible if we set α such that $0 < \alpha < 0.4$.

The next step is to give the asymptotic variance-covariance of the first term, that is,

$$\sqrt{m_n} \left[\frac{1}{m_n} \sum_{k=1}^{m_n} \mathbf{x}_k^* \mathbf{x}_k^{*'} - \boldsymbol{\Sigma}_x \right] \quad (62)$$

because it is of the order $O_p(1)$.

Let the return vector in $(t_{i-1}^{(n)}, t_i^{(n)})$ be $\mathbf{r}_i^{(n)} = \mathbf{X}(t_i^{(n)}) - \mathbf{X}(t_{i-1}^{(n)}) = (r_{ig}^{(n)})$ ($i = 1, \dots, n; g = 1, \dots, p$). Then, we can write

$$\begin{aligned} & \frac{1}{m_n} \sum_{k=1}^{m_n} \mathbf{x}_k^* \mathbf{x}_k^{*'} \\ &= \frac{1}{m_n} \left(\frac{2n}{n + \frac{1}{2}} \right) \sum_{k=1}^{m_n} \left[\sum_{i=1}^n \mathbf{r}_i \cos \left[\pi \left(\frac{2k-1}{2n+1} \right) \left(i - \frac{1}{2} \right) \right] \sum_{j=1}^n \mathbf{r}_j' \cos \left[\pi \left(\frac{2k-1}{2n+1} \right) \left(j - \frac{1}{2} \right) \right] \right] \\ &= \sum_{i=1}^n c_{ii} \mathbf{r}_i \mathbf{r}_i' + \sum_{i \neq j} c_{ij} \mathbf{r}_i \mathbf{r}_j', \end{aligned}$$

where

$$\begin{aligned} c_{ii} &= \left(\frac{2n}{2n+1} \right) \left[1 + \frac{1}{m} \frac{\sin 2\pi m \left(\frac{i-1/2}{2n+1} \right)}{\sin \left(\pi \frac{i-1/2}{2n+1} \right)} \right], \\ c_{ij} &= \frac{1}{2m} \left(\frac{2n}{2n+1} \right) \left[\frac{\sin 2\pi m \left(\frac{i+j-1}{2n+1} \right)}{\sin \left(\pi \frac{i+j-1}{2n+1} \right)} + \frac{\sin 2\pi m \left(\frac{j-i}{2n+1} \right)}{\sin \left(\pi \frac{j-i}{2n+1} \right)} \right] \quad (i \neq j). \end{aligned}$$

Then, KSK (2018) have shown that $\sqrt{m_n} \sum_{i=1}^n \left[\mathbf{r}_i^{(n)} \mathbf{r}_i^{(n)'} - \boldsymbol{\Sigma}_x + (c_{ii} - 1) \mathbf{r}_i^{(n)} \mathbf{r}_i^{(n)'} \right] = o_p(1)$ and (62) can be written as $\sqrt{m_n} \sum_{i=1}^n \left[c_{ii} \mathbf{r}_i^{(n)} \mathbf{r}_i^{(n)'} - \boldsymbol{\Sigma}_x \right] + \frac{\sqrt{m_n}}{n} \sum_{i \neq j}^n \left[c_{ij} \mathbf{r}_i^{(m)} \mathbf{r}_j^{(n)'} \right]$.

We need to evaluate the asymptotic variance of its second term, which consists of off-diagonal parts, because the first term, which consists of diagonal terms, is stochastically negligible. KSK (2018) have also shown that when there is no jump term the variance of the limiting distribution of the (g, h) th element of the limiting variance-covariance matrix is the limit of

$$V_n(g, h) = m_n \sum_{i,j=1}^n c_{ij}^2 \left[\int_{t_{i-1}^{(n)}}^{t_i^{(n)}} \mathbf{c}_{gg}(s) ds \int_{t_{j-1}^{(n)}}^{t_j^{(n)}} \mathbf{c}_{hh}(s) ds + \int_{t_{i-1}^{(n)}}^{t_i^{(n)}} \mathbf{c}_{gh}(s) ds \int_{t_{j-1}^{(n)}}^{t_j^{(n)}} \mathbf{c}_{gh}(s) ds \right].$$

By using the relation

$$\sum_{i,j=1}^n c_{ij}^2 = \frac{4}{m_n} \left[\frac{n}{2} + \frac{1}{4} \right]^2 \quad (63)$$

KSK (2018) have shown in Lemma 5.6 that

$$V_n(g, h) \xrightarrow{p} V(g, h) = \int_0^1 [\mathbf{c}_{gg}(s)\mathbf{c}_{hh}(s) + \mathbf{c}_{gh}^2(s)]ds \quad (64)$$

$$U_n(g, h) = \sum_{j=2}^n \left[2 \sum_{i=1}^{j-1} \sqrt{m_n} c_{ij} r_{ig}^{(n)} \right] r_{jh}^{(n)}, \quad (65)$$

where $c_{ij} = (2/m) \sum_{k=1}^m s_{ik}s_{jk}$ ($i, j = 1, \dots, n$) and $s_{jk} = \cos \left[\frac{2\pi}{2n+1} (j - \frac{1}{2})(k - \frac{1}{2}) \right]$.

Also the asymptotic covariance of (g, h) th element of $\hat{\Sigma}_x$ and (k, l) th element of $\hat{\Sigma}_x$ for $g, h, k, l = 1, \dots, p$ can be calculated as

$$CV(g, h; k, l) = \int_0^1 [\mathbf{c}_{gk}(s)\mathbf{c}_{hl}(s) + \mathbf{c}_{gl}(s)\mathbf{c}_{hk}(s)]ds.$$

(The actual calculations are complicated, but the resulting form of covariance has been quite common in statistical multivariate analysis. See Chapter 3 of Anderson (2003), for instance.)

By using the fact that the fourth-order moments of martingales are of smaller orders and negligible asymptotically as $n \rightarrow \infty$ (see Chapter 5 of KSK (2018)), we have the asymptotic normality with the covariance matrix $V(g, h)$ if it is deterministic.

When $V(g, h)$ is random, we need to use the conditional variance and covariances of martingales. The expression and derivation becomes lengthy, but it may be straightforward. In addition to this calculations, we need some extra arguments in the following Step 3 for the stable convergence of the limiting random variables in a proper mathematical sense.

(Step 2) We can extend Step 1 to the cases with jump terms. For simplicity, we consider the case when the number of jumps is finite and the jump sizes are bounded. For $g = 1, \dots, p$, let $\sigma(t) = (\sigma_g(t))$ and

$$\begin{aligned} r_{ig}^{(n)} &= \sigma_g(t_{i-1}^{(n)})[\mathbf{W}(t_i^{(n)}) - \mathbf{W}(t_{i-1}^{(n)})] + \sum_{t_{i-1}^{(n)} \leq s \leq t_i^{(n)}} \Delta X_g(s) \\ &= R_{ig}^{(1)} + R_{ig}^{(2)} \text{ (say)} \end{aligned} \quad (66)$$

We decompose

$$\sum_{i,j=1}^n c_{ij} \left[r_{ig}^{(n)} r_{jh}^{(n)} - R_{ig}^{(2)} R_{jh}^{(2)} - \delta(i, j) \left(\int_{t_{i-1}^{(n)}}^{t_i^{(n)}} c_{gh}(s) ds \right) \right]$$

into two parts (i) + (ii), where (i) is the summation with $\sum_{i=j=1}^n$ and (ii) is the summation with $\sum_{i \neq j}$. The effects of $\sum_{i \neq j=1}^n c_{ij} R_{ig}^{(2)} R_{jh}^{(2)} = o_p(\frac{1}{m})$ as $m \rightarrow \infty$ because

$$\begin{aligned}
c_{ij} &= \frac{2}{m} \sum_{k=1}^m s_{ik} s_{jk} \\
&= \frac{1}{m} \sum_{k=1}^m \cos \left[\frac{2\pi}{2n+1} (i+j-1) \left(k - \frac{1}{2} \right) \right] \cos \left[\frac{2\pi}{2n+1} (i-j) \left(k - \frac{1}{2} \right) \right],
\end{aligned}$$

and the assumptions of boundedness and finiteness of jumps. (They can be relaxed with some additional complications.)

Moreover, $\sqrt{m} \sum_{i \neq j=1}^n c_{ij} R_{ig}^{(2)} R_{jh}^{(2)} = o_p\left(\frac{1}{\sqrt{m}}\right)$ as $m \rightarrow \infty$ for the asymptotic distribution.

As in Step 1, we write (i)

$$\begin{aligned}
&\sum_{i=j=1}^n \left[r_{ig}^{(n)} r_{jh}^{(n)} - \left(\int_{t_{i-1}^{(n)}}^{t_i^{(n)}} c_{gh}(s) ds + R_{ig}^{(2)} R_{ih}^{(2)} \right) \right. \\
&\quad \left. + (c_{ii} - 1) \left(r_{ig}^{(n)} r_{jh}^{(n)} - \int_{t_{i-1}^{(n)}}^{t_i^{(n)}} c_{gh}(s) ds - R_{ig}^{(2)} R_{jh}^{(2)} \right) \right].
\end{aligned}$$

Since

$$c_{jj} - 1 = \frac{1}{2m} \frac{\sin \left[\frac{2\pi m}{2n+1} (2j-1) \right]}{\sin \left[\frac{\pi}{2n+1} (2j-1) \right]} \quad (j = 1, \dots, m) \quad (67)$$

by using an extension of Lemma 5.4 in KSK (2018),

$$\sqrt{m} \sum_{i=j=1}^n \left[(c_{ii} - 1) \left(r_{ig}^{(n)} r_{jh}^{(n)} - \int_{t_{i-1}^{(n)}}^{t_i^{(n)}} c_{gh}(s) ds - R_{ig}^{(2)} R_{jh}^{(2)} \right) \right]$$

is negligible when $m, n \rightarrow \infty$ and $m/n \rightarrow 0$, and

$$\sum_{i=1}^n \left[r_{ig}^{(n)} r_{ih}^{(n)} - \left(\int_{t_{i-1}^{(n)}}^{t_i^{(n)}} c_{gh}(s) ds + \sum_{t_{i-1}^{(n)} \leq s \leq t_i^{(n)}} \Delta X_g(s) \Delta X_h(s) \right) \right] \xrightarrow{p} 0 \quad (68)$$

as $m \rightarrow \infty$.

The evaluation of the asymptotic distribution can be done as in Step 1, but there are some additional complication. The main terms of $\mathcal{F}_0$ -conditional covariance with jumps can be calculated from the off-diagonal parts as

$$m_n \sum_{i \neq j} c_{ij}^2 \{ [(R_{ig}^{(1)} + R_{ig}^{(2)})^2 (R_{jh}^{(1)} + R_{jh}^{(2)})^2] - [(R_{ig}^{(1)})^2 (R_{jh}^{(1)})^2 + (R_{ig}^{(2)})^2 (R_{jh}^{(2)})^2] \}$$

By evaluating the orders of each terms, we find that the important terms are

$$m_n \sum_{i \neq j} c_{ij}^2 \{[(R_{ig}^{(1)})^2 (R_{jh}^{(2)})^2 + (R_{jh}^{(1)})^2 (R_{ig}^{(2)})^2] + [(R_{ig}^{(1)})(R_{ih}^{(1)})(R_{jg}^{(2)})(R_{jh}^{(2)}) \\ + (R_{ig}^{(2)})(R_{ih}^{(2)})(R_{jg}^{(1)})(R_{jh}^{(1)})]\}$$

and other terms are of the negligible orders. The effects of diagonal elements in the summation $\sum_{i=j=1}^n [\cdot]$ are of smaller orders asymptotically as $m_n, n \rightarrow \infty$ and $m_n/n \rightarrow 0$, but it is appropriate to add them to the terms from off-diagonal elements to obtain the resulting simple form. Then, we can derive V_{gh} in Theorem 2.1.

When there can be jumps, $U_n(g, h)$ in Step 1 should be slightly changed as $U_n(g, h) = \sum_{j=2}^n 2 \sum_{i=1}^{j-1} \sqrt{m_n} c_{ij} [r_{ig}^{(n)} r_{jh}^{(n)} - R_{ig}^{(2)} R_{jh}^{(2)}]$.

By using the fact that the fourth-order moments of martingales are negligible, we have the asymptotic normality when the covariance matrix is deterministic. When $V(g, h)$ is random, we need to use the conditional variance and covariances of martingales. (The derivation is similar to the case without jumps.) In addition to this calculations, we need some extra arguments in the following Step 3 for the stable convergence of the limiting random variables.

We omit the derivation of $V_{gh,kl}$, which is a result of lengthy calculations, but the result is similar to the corresponding one in Step 1.

(Step 3) We need the martingale CLT to prove the asymptotic normality of the SIML estimator. Also some conditions are needed for the stable convergences of asymptotic distribution because the asymptotic variance-covariances are random in the general cases. [See Theorem 2.2.15 of Jacod and Protter (2012), for instance.] We only outline some of important steps, and we set $p = q_1 = q_3 = 1$ and $q_2 = 0$ in Step 3. (The generalization may be straightforward.)

Let

$$S_m = \sqrt{m} \sum_{j>i=1}^n c_{ij} r_i^{(n)} r_j^{(n)}, \quad (69)$$

$$\text{and} \quad M_{1n} = \sum_{j=1}^n r_{jg}^{(n)}, \quad M_{2n} = \sum_{j=1}^n (W(t_j)^\sigma - W(t_{j-1})^\sigma) \quad \text{and} \\ r_j^{(n)} = \int_{t_{j-1}^{(n)}}^{t_j^{(n)}} \sigma_s dW(s) \quad (j = 1, \dots, n).$$

We need to show that S_n and M_{in} ($i = 1, 2$) are orthogonal. But then we shall show

$$\sum_{j=2}^n \mathbf{E} \left[\left(\sqrt{m} \sum_{i=1}^{j-1} c_{ij} r_i^{(n)} \right) r_j^{(n)} r_j^{(n)} | \mathcal{F}_{j-1} \right] = \sum_{j=2}^n \left[\int_{t_{j-1}^{(n)}}^{t_j^{(n)}} \sigma_s ds \right] \left[\left(\sqrt{m} \sum_{i=1}^{j-1} c_{ij} r_i^{(n)} \right) \right]^p \rightarrow 0$$

as $n \rightarrow \infty$.

Assume that σ_s ($t_{j-1}^{(n)} \leq s \leq t_j^{(n)}$) is bounded and orthogonal to $r_i^{(n)} = \int_{t_{i-1}^{(n)}}^{t_i^{(n)}} \sigma_s dW(s)$ for $i < j$. Then, it suffices to prove

$$\text{Var} \left[\frac{1}{n} \sum_{i=1}^{n-1} \sqrt{m} \left(\sum_{j=i+1}^n c_{ij} \right) r_i^{(n)} \right] \sim \left(\frac{m}{n} \right) \sum_{i=1}^{n-1} \left[\frac{1}{n} \sum_{j=i+1}^n c_{ij} \right]^2 \rightarrow 0$$

as $n \rightarrow \infty$. After some calculations, we find that

$$\begin{aligned} \frac{1}{n} \sum_{j=i+1}^n c_{ij} &= \frac{1}{m} \sum_{k=1}^m \cos \frac{2\pi}{2n+1} \left(i - \frac{1}{2} \right) \left(k - \frac{1}{2} \right) \\ &\quad \left[\frac{-\sin \frac{2\pi}{2n+1} i \left(k - \frac{1}{2} \right) + \sin \frac{2\pi}{2n+1} n \left(k - \frac{1}{2} \right)}{n \sin \frac{2\pi}{2n+1} \frac{1}{2} \left(k - \frac{1}{2} \right)} \right], \end{aligned}$$

which is bounded.

We have the last evaluation because we can use $c_{ij} = (2/m) \sum_{k=1}^m s_{ik} s_{jk}$, $s_{jk} = \cos[\frac{2\pi}{2n+1} (j - \frac{1}{2})(k - \frac{1}{2})]$. Then we have

$$\begin{aligned} &\sum_{j=i+1}^n 2 \cos \left[\frac{2\pi}{2n+1} \left(j - \frac{1}{2} \right) \left(k - \frac{1}{2} \right) \right] \\ &= \sum_{j=i+1}^n \left[e^{i \frac{2\pi}{2n+1} \left(j - \frac{1}{2} \right) \left(k - \frac{1}{2} \right)} + e^{-i \frac{2\pi}{2n+1} \left(j - \frac{1}{2} \right) \left(k - \frac{1}{2} \right)} \right] \\ &= e^{i \frac{2\pi}{2n+1} \left(i + \frac{1}{2} \right) \left(k - \frac{1}{2} \right)} \frac{1 - e^{i \frac{2\pi}{2n+1} (n-i) \left(k - \frac{1}{2} \right)}}{1 - e^{i \frac{2\pi}{2n+1} \left(k - \frac{1}{2} \right)}} + e^{-i \frac{2\pi}{2n+1} \left(i + \frac{1}{2} \right) \left(k - \frac{1}{2} \right)} \frac{1 - e^{-i \frac{2\pi}{2n+1} (n-i) \left(k - \frac{1}{2} \right)}}{1 - e^{-i \frac{2\pi}{2n+1} \left(k - \frac{1}{2} \right)}} \end{aligned}$$

and we can evaluate the summations as in Chapter 5 of KSK (2018).

We also have the asymptotic orthogonality between S_m and M_{1n} (or $M_{1n}^* = \sum_{j=1}^n (W(t_j) - W(t_{j-1}))$, or M_{2n}).

Then, it can be shown that the conditional covariance terms such as $\int_0^1 c_{gg}(s) c_{hh}(s) ds$ are asymptotically orthogonal to S_m . It is because by using the assumptions on the volatility process of (3), it has the solution in an integral form with bounded coefficients (Chapter IV of Ikeda and Watanabe (1989)). Then, we can use the discretized version of $c_{gg}(s) c_{hh}(s)$ in any interval $(t_{j-1}^n, t_j^n]$ ($j = 1, \dots, n$) with $W(t_j)^\sigma - W(t_{j-1})^\sigma$ to show an asymptotic orthogonality.

Finally, the weak convergence and the stable CLT are used for the normalized SIML estimator when the underlying process is the Ito' semi-martingale processes containing jump terms. The martingale part of jumps can be written as

$$M_{3n} = \sum_{j=1}^n \left[\sum_{t_{j-1}^{(n)} < s \leq t_j^{(n)}} \Delta X(s) \right]. \quad (70)$$

Here, we assume that there is no natural increasing process parts in jumps. (See Ikeda and Watanabe (1989) for instance). The effects of increasing process parts are stochastically negligible under the present assumption of finiteness and boundedness

of jumps because the number of large jumps in $[0, 1]$ is finite. Then, it is possible to find that S_m and M_{3n} are orthogonal asymptotically because each element of (3) contains only Brownian motions and the product of Brownian motions and jump terms is independent. $\square$

(ii) Derivation of Lemma 3.2

We consider the normalized random variable $\mathbf{b}' \sqrt{\frac{a}{a_m}} \mathbf{g}_m$ for a (nonzero) constant vector $\mathbf{b}$. Because of independence of $\mathbf{X}(t)$ and $\mathbf{v}(t)$, the variance can be written as:

$$\sum_{k=1}^m \mathcal{E}[\mathbf{b}' \mathbf{x}_k^* \mathbf{x}_k^{*'} \mathbf{b}] \left[\frac{a_{kn}}{a_m} \right] \boldsymbol{\beta}_j' \boldsymbol{\Sigma}_v \boldsymbol{\beta}_j. \quad (71)$$

We write the (true) return vector process $\mathbf{r}_i^{(n)} = \mathbf{X}(t_i^n) - \mathbf{X}(t_{i-1}^n) (= (r_{ig}^{(n)})$ ($i = 1, \dots, n; g = 1, \dots, p$) and $\mathbf{x}_k^* = \sqrt{\frac{4n}{2n+1}} \sum_{j=1}^n \mathbf{r}_j^{(n)} \cos[\pi(2k-1)/(2n+1)(j-1/2)]$. Then, by direct calculation, we find that

$$\mathcal{E}[\mathbf{r}_i^{(n)} \mathbf{r}_i^{(n)'}] - \mathcal{E} \left[\int_{t_{i-1}^n}^{t_i^n} \mathbf{c}(s) ds + \sum_{t_{i-1}^n \leq s < t_i^n} (\Delta \mathbf{X}(s)) (\Delta \mathbf{X}(s))' \right] = o\left(\frac{1}{n}\right).$$

Because of Assumption 1, Lemma 3.1 and the boundedness assumption of $\boldsymbol{\sigma}(s)$ and $\Delta \mathbf{X}(s)$ ($0 \leq s \leq 1$), the variance of the normalized $\mathbf{g}_m$ is positive and bounded.

(iii) Derivation of Lemma 3.3

We use the CLT for a martingale, which is Q_2 below. (In this case, the asymptotic variance of a limiting variable is constant and it is different from Theorem 2.1.)

For this purpose, we set the transformation $\mathbf{K}_n = \sqrt{n} \mathbf{P}_n \mathbf{C}_n^{-1} = (b_{kj})$ and then, we have $\sum_{j=1}^n b_{kj} b_{k'j} = \delta(k, k') a_{kn}$ and $a_{kn} = 4n \sin^2[(\pi/2)(2k-1)/(2n+1)]$ ($k = 1, \dots, n$).

Let $\Gamma_{lj} = \mathbf{v}_j' \boldsymbol{\beta}_l$ for a constant (nonzero) vector $\boldsymbol{\beta}_l$ ($l = 1, \dots, p$)

$$Q = \sum_{k=1}^m \left[\left(\sum_{j=1}^n b_{kj} \Gamma_{lj} \sum_{j'=1}^n b_{kj'} \Gamma_{l'j'} - a_{kn} \boldsymbol{\beta}_l' \boldsymbol{\Sigma}_v \boldsymbol{\beta}_{l'} \right) \right]. \quad (72)$$

We decompose Q into two parts $Q = Q_1 + Q_2$ and write

$$Q_1 = \sum_{k=1}^m \left[\sum_{j=j'=1}^n b_{kj}^2 \Gamma_{lj} \Gamma_{l'j} - a_{kn} \boldsymbol{\beta}_l' \boldsymbol{\Sigma}_v \boldsymbol{\beta}_{l'} \right]. \quad (73)$$

and

$$Q_2 = \sum_{k=1}^m \left[\sum_{j \neq j'=1}^n b_{kj} b_{kj'} \Gamma_{lj} \Gamma_{l'j'} \right]. \quad (74)$$

The dominant term in Q is Q_2 stochastically, which consists of off-diagonal elements of Q and they can be rewritten as martingales. The term Q_1 consists of diagonal elements, and the stochastic order is smaller than Q_2 . (This structure is similar to (62). Hence, we use a martingale CLT for the normalized Q_2 . The variance of the dominant term becomes

$$\begin{aligned} & \mathbf{E} \left[\sum_{k=1}^m \sum_{j \neq j'=1}^n b_{kj} b_{k,j'} \Gamma_{lj} \Gamma_{l'j'} \right]^2 \\ &= \left[\sum_{k,k'=1}^m \sum_{j \neq j'=1}^n b_{kj} b_{k,j'} b_{k'j} b_{k',j'} [\mathbf{E}(\Gamma_{lj}^2) \mathbf{E}(\Gamma_{l'j'}^2) + (\mathbf{E}(\Gamma_{lj} \Gamma_{l'j'})^2] \right] \\ &= \left[\sum_{k,k'=1}^m \sum_{j,j'=1}^n b_{kj} b_{k'j} b_{k'j'} b_{kj'} \delta(k, k') \mathbf{E}(\Gamma_{lj}^2) \mathbf{E}(\Gamma_{l'j'}^2) + (\mathbf{E}(\Gamma_{lj} \Gamma_{l'j'})^2 \right] \\ &\quad - 2 \left[\sum_{k,k'=1}^m \sum_{j=1}^n b_{kj}^2 b_{k'j}^2 \right] [\mathbf{E}(\Gamma_{lj}^2)]^2. \end{aligned}$$

The first term is dominant as we shall show and it becomes

$$\sum_{k=k'=1}^m a_{kn}^2 [\mathbf{E}(\Gamma_l^2) \mathbf{E}(\Gamma_{l'}^2) + (\mathbf{E}(\Gamma_l \Gamma_{l'})^2] = m \times a_2(m) [\mathbf{E}(\Gamma_l^2) \mathbf{E}(\Gamma_{l'}^2) + (\mathbf{E}(\Gamma_l \Gamma_{l'})^2],$$

which is $O(\frac{m^5}{n^2})$ because $a_m(2) = O(m^2/n^2)$ (see Lemma 3.1). The asymptotic variance of the normalized Q_2 is given by

$$\mathbf{A}\mathbf{V} \left[\frac{1}{a_m \sqrt{m}} Q_2 \right] = \left[\lim_{m \rightarrow \infty} \frac{a_m(2)}{a_m^2} \right] [\mathbf{E}(\Gamma_l^2) \mathbf{E}(\Gamma_{l'}^2) + (\mathbf{E}(\Gamma_l \Gamma_{l'})^2)]. \quad (75)$$

We denote $Q_2(l, l')$ as the above Q_2 and we set $Q_2(h, h')$ as another Q_2 with $\Gamma_{hj} = \mathbf{v}'_j \boldsymbol{\beta}_h$, and $\Gamma_{h'j} = \mathbf{v}'_j \boldsymbol{\beta}_{h'}$ ($l, l', h, h' = 1, \dots, p; j = 1, \dots, n$). By using similar evaluations, we find that the asymptotic covariance of $Q_2(l, l')$ and $Q_2(h, h')$ is given as

$$\left[\lim_{m \rightarrow \infty} \frac{a_m(2)}{a_m^2} \right] [\mathbf{E}(\Gamma_l \Gamma_h) \mathbf{E}(\Gamma_{l'} \Gamma_{h'}) + \mathbf{E}(\Gamma_l \Gamma_{h'}) \mathbf{E}(\Gamma_{l'} \Gamma_h)]. \quad (76)$$

From (73), it is possible to show that the stochastic order of Q_1 is smaller than the order of Q_2 because the following lemma shows that

$$\mathbf{E}[(Q_1)^2] = O \left(\left[\sum_{k,k'=1}^m \left(\sum_{j=1}^n b_{kj}^2 b_{k'j}^2 \right) \right] \right) \quad (77)$$

under the assumption of $\mathbf{E}[\|\mathbf{v}_j\|^2] < \infty$. Since the normalizing factor is $1/(a_m \sqrt{m})$ and $a_m = O(m^2/n)$ (Lemma 3.1), we have

$$\left[\frac{1}{a_m \sqrt{m}} \right]^2 \times \left[\sum_{k,k'=1}^m \left(\sum_{j=1}^n b_{kj}^2 b_{k'j}^2 \right) \right] = O\left(\frac{1}{mn}\right).$$

We prepare a key lemma.

$$\sum_{j=1}^n b_{kj}^2 b_{k'j}^2 = O\left(n \times \sin^2 \frac{\theta_k}{2} \times \sin^2 \frac{\theta_{k'}}{2}\right) \quad (78)$$

Lemma A.1

and

$$\sum_{k,k'=1}^m \sum_{j=1}^n b_{kj}^2 b_{k'j}^2 = o(m \times a_m(2)), \quad (79)$$

where $\theta_k = \frac{2\pi}{2n+1}(k - \frac{1}{2})$.

Proof of Lemma A.1 We use the evaluation method of proof of Lemma A-1 in Kunitomo and Sato (2017). We set $b_{kj} = p_{kj} - p_{k,j+s}$ ($1 \leq j \leq n-1$) and

$$b_{kj} = \frac{\sqrt{n}}{\sqrt{2n+1}} \left\{ \left[1 - e^{i \frac{2\pi}{2n+1} \left(k - \frac{1}{2}\right)s} \right] e^{i \frac{2\pi}{2n+1} \left(k - \frac{1}{2}\right) \left(j - \frac{1}{2}\right)} + \left[1 - e^{-i \frac{2\pi}{2n+1} \left(k - \frac{1}{2}\right)s} \right] e^{-i \frac{2\pi}{2n+1} \left(k - \frac{1}{2}\right) \left(j - \frac{1}{2}\right)} \right\}.$$

Then, we need to evaluate each term of

$$\sum_{j=1}^{n-1} [b_{kj}]^2 [b_{k'j}]^2 = \frac{n}{2n+1} \sum_{j=1}^{n-1} \{ [A_j(k) + \bar{A}_j(k)]^2 [A_j(k') + \bar{A}_j(k')]^2 \},$$

where $A_j(k) = 1 - e^{-i \frac{2\pi}{2n+1} \left(k - \frac{1}{2}\right)s}$ and $\bar{A}_j$ is the complex conjugate of $A_j(k)$.

There are 9 terms, but the typical term, for instance, becomes

$$\begin{aligned} & \sum_{j=1}^{n-1} A_j(k)^2 A_j(k')^2 \\ &= (1 - e^{i\theta_k^s})^2 (1 - e^{i\theta_{k'}^s})^2 \frac{1 - e^{i \frac{2\pi}{2n+1} 2(k+k'-1)(n-1+1)}}{1 - e^{i \frac{2\pi}{2n+1} 2(k+k'-1)}} \times e^{i \frac{2\pi}{2n+1} 2(k+k'-1) \frac{1}{2}} \\ &= (e^{-i \frac{\theta_k^s}{2}} - e^{i \frac{\theta_k^s}{2}})^2 (e^{-i \frac{\theta_{k'}^s}{2}} - e^{i \frac{\theta_{k'}^s}{2}})^2 \frac{e^{-i \frac{2\pi}{2n+1} (k+k'-1)} - e^{i \frac{2\pi}{2n+1} (k+k'-1)(2n+1-1/2)}}{e^{-i \frac{2\pi}{2n+1} (k+k'-1)} - e^{i \frac{2\pi}{2n+1} (k+k'-1)}}. \end{aligned}$$

Hence, for $k = 1, \dots, m$ and $m/n \rightarrow 0$ ($n \rightarrow \infty$), we can find a positive constant c_1 such that

$$\left| \sum_{j=1}^{n-1} A_j(k)^2 A_j(k')^2 \right| \leq c_1 \times n \times \sin^2 \left[\frac{\theta_k}{2} \right] \times \sin^2 \left[\frac{\theta_{k'}}{2} \right].$$

Then, after lengthy but straightforward calculations to each terms, we find that

$$\sum_{k,k'=1}^m \sum_{j=1}^n b_{kj}^2 b_{k'j}^2 = O(n \times (m \times a_m)^2) = O\left(n \times \left(\frac{1}{n} \frac{m^2}{n}\right)^2\right), \quad (80)$$

which is $o(m \times a_m(2))$, provided that $m = [n^\alpha]$ ($0 < \alpha < 1$). $\square$

Lemma A.2 *The normalized Q_2 has a martingale representation*

$$Q_2^{(n)} = \sum_{j=2}^n \left[2 \sum_{j'=1}^{j-1} \left(\sum_{k=1}^m b_{kj} b_{kj'} \right) \Gamma_{l'j'} \right] \Gamma_{lj} \quad (81)$$

and the asymptotic distribution is normal as $n \rightarrow \infty$ under the condition of $\mathbf{E}[\|\mathbf{v}_i\|^4] < \infty$.

The asymptotic covariance of the limiting random variable is given by (75).

Proof of Lemma A.2 We follow the evaluation method in the first two steps of the proof of Theorem 2.1. Since Q_2 is a martingale and $\mathbf{v}_i$ are sequences of independent random variables, we can use the standard martingale CLT. (The variance and covariances of the limiting random variables are constant in the present case.)

The asymptotic variance has been calculated in (75). The sufficient condition can be checked by the calculation of the fourth-order moments of Q_2 . The evaluations may look complicated, but they are straightforward because of the assumption of the existence of fourth order moments of independent noise terms. We illustrate some of derivations of

$$\mathbf{E} \left[\left(\frac{Q_2}{2} \right)^4 \right] = \mathbf{E} \left[\sum_{j=2}^n \left[\sum_{j'=1}^{j-1} \left(\sum_{k=1}^m b_{kj} b_{kj'} \right) \Gamma_{l'j'} \right] \Gamma_{lj} \right]^4. \quad (82)$$

There are many terms, but it is enough to evaluate the orders of dominant terms, that is, a constant times

$$\sum_{j,j'=2}^n \sum_{i,i'=1}^n \left[\sum_{k_1,k_2,k_3,k_4=1}^m (b_{k_1j} b_{k_2j} b_{k_1i} b_{k_2i}) (b_{k_3j'} b_{k_4j'} b_{k_3i'} b_{k_4i'}) \right].$$

By using Lemma 3.1, it becomes

$$\left(\sum_{k_1=k_2=1}^m a_{k_1,n}^2 \right) \left(\sum_{k_3=k_4=1}^m a_{k_3,n}^2 \right) = o\left(\left(\frac{m^4}{n^2} \right)^2 \right).$$

Since the normalization term is $(1/(a_m \sqrt{m}))^4 = n^4/m^{10}$, the fourth-order moment of the normalized Q_2 is $O(1/m^2)$, which converges to 0 as $m \rightarrow \infty$ when $n \rightarrow \infty$. $\square$

(iv) **Derivation of Theorem 3.4**

(iv-1) : Let $A_{r,n} = \{T_n(r) \geq \chi_{1-\alpha_n}^2(1)\}$, $r = 1, \dots, p-1$. Note that $P(\hat{q}_x^* = p - (r_0 - 1)) = P(\cap_{j=r_0}^{p-1} A_{j,n} \cap A_{r_0-1,n}^c)$. Since $\alpha_n = \alpha/\log m_n \rightarrow 0$ and $-\log \alpha_n = o(m/a_m(2))$, by Theorem 5.8 in Pötscher (1983), we have that $(m/a_m)(2)^{-1} \chi_{1-\alpha_n}^2(1) \rightarrow 0$ as $n \rightarrow \infty$. If $p - (r_0 - 1) < q_x^*$, then we have that as $n \rightarrow \infty$,

$$\begin{aligned} P(\hat{q}_x = p - (r_0 - 1)) &\leq P(A_{r_0-1,n}^c) = 1 - P(A_{r_0-1,n}) \\ &= 1 - P\left((m/a_m(2))^{-1} T_n(r_0 - 1) \geq (m/a_m(2))^{-1} \chi_{1-\alpha_n}^2(1)\right) \rightarrow 0. \end{aligned}$$

Indeed, we can show that $T_n(r_0 - 1) = O_p(m/a_m(2))$ if $p - (r_0 - 1) < q_x^*$. If $p - (r_0 - 1) > q_x^*$, then Theorem 3.3 yields that

$$P(\hat{q}_x = p - (r_0 - 1)) \leq P(A_{r_x^*,n}) = P(T_n(r_x^*) \geq \chi_{1-\alpha_n}^2(1)) \rightarrow 0$$

as $n \rightarrow \infty$. Therefore, we obtain $\hat{q}_x^* \xrightarrow{p} q_x^*$.

(iv-2) : When $q_x^* = 0$ or p , we have the same result.

(v) Finally, for the sake of convenience, we summarize an important, but simple calculation of variances used in Theorem 3.3, Corollaries 3.1 and 3.2 as a lemma.

Lemma A.3 Assume the normality of market microstructure noises. For any $p \times p$ fixed positive definite matrix $\mathbf{H}$, let a $p \times p$ matrix

$$\mathbf{U}^* = (U_{ij}^*) = (\mathbf{B}' \mathbf{H} \mathbf{B})^{-1} \left(\sqrt{\frac{m}{a_m(2)}} \right) \mathbf{B}' (\mathbf{G}_m - \boldsymbol{\Sigma}_m) \mathbf{B}, \quad (83)$$

and $\mathbf{A}^{-1} = (a^{ij}) = (\mathbf{B}' \mathbf{H} \mathbf{B})^{-1}$ and $\boldsymbol{\Omega}_v = (\omega_{ij}) = (\mathbf{B}' \boldsymbol{\Sigma}_v \mathbf{B})$. We denote X^* as the limiting random matrix of $\sqrt{\frac{m}{a_m(2)}} \mathbf{B}' (\mathbf{G}_m - \boldsymbol{\Sigma}_m) \mathbf{B}$, which is proportional to the limiting random variables $\mathbf{E}_m^* = (1/\sqrt{m}) \sum_{k=1}^m a_{kn} (\mathbf{u}_k^* \mathbf{u}_k^{*'} - \boldsymbol{\Omega}_v)$. Then, the asymptotic variance is given by

$$\begin{aligned}
\mathbf{AV}(U_{ii}^*) &= \mathcal{E} \left[\sum_{j=1}^p a^{ij} X_{ji}^* \sum_{j'=1}^p a^{ij'} X_{j'i}^* \right] \\
&= \sum_{j,j'=1}^p a^{ij} a^{ij'} \text{Cov}(X_{ji}^*, X_{j'i}^*) \\
&= \sum_{j,j'=1}^p a^{ij} a^{ij'} [\omega_{jj'} \omega_{ii} + \omega_{ji} \omega_{j'i}] .
\end{aligned} \tag{84}$$

When $\mathbf{H} = \Sigma_v$, it becomes

$$\omega_{ii} \sum_{j,j'=1}^p \omega^{ij} \omega^{ij'} \omega_{jj'} + \sum_{j,j'=1}^p \omega^{ij} \omega_{ji} \omega^{ij'} \omega_{j'i} = \omega_{ii} \omega^{ii} + 1 . \tag{85}$$

Also we find the asymptotic covariances as

$$\begin{aligned}
\mathbf{ACov}(U_{ii}^*, U_{kk}^*) &= \mathcal{E} \left[\sum_{j=1}^p a^{ij} X_{ji}^* \sum_{j'=1}^p a^{kj'} X_{j'k}^* \right] \\
&= \sum_{j,j'=1}^p a^{ij} a^{kj'} \text{Cov}(X_{ji}^*, X_{j'k}^*) \\
&= \sum_{j,j'=1}^p a^{ij} a^{kj'} [\omega_{jj'} \omega_{ik} + \omega_{jk} \omega_{j'i}] \\
&= \sum_{j'=1}^p \left[\sum_{j=1}^p a^{ij} \omega_{jj'} \right] a^{kj'} \omega_{ik} + \sum_{j,j'=1}^p a^{ij} \omega_{jk} a^{kj'} \omega_{j'i} .
\end{aligned}$$

When $\mathbf{H} = \Omega_v$, it becomes

$$\sum_{j'=1}^p \omega^{kj'} \omega_{ik} + 1 = \omega^{ki} \omega_{ik} + 1 . \tag{86}$$

Acknowledgements This is a revision of MIMS-RBP Statistics and Data Science Series (SDS-10), Meiji University. We thank seminar participants at International Symposium on Econometric Theory and Applications at Osaka for helpful comments and Hiroumi Misaki (Tsukuba University) for giving useful information on high-frequency financial data at JXP. We also thank anonymous referees for their comments to the submitted version. The research was supported by Grant-in-Aid for Scientific Research (B) (17H02513) from JSPS, and Daisuke Kurisu is supported by Grant-in-Aid for Research Activity Start-up (19K20881) and Early Career Scientists (20K13468) from JSPS.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the

material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aït-Sahalia, Y., & Jacod, J. (2014). *High-frequency financial econometrics*. Princeton: Princeton University Press.
- Aït-Sahalia, Y., & Xiu, D. (2019). Principal component analysis of high frequency data. *Journal of the American Statistical Association*, 114, 287–303.
- Anderson, T. W. (1984). Estimating linear statistical relationships. *The Annals of Statistics*, 12, 1–45.
- Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley.
- Bibinger, M., Hautsch, N., Malec, P., & Reiß, M. (2014). Estimating the quadratic covariance matrix from noisy observations : Local method of moments and efficiency. *The Annals of Statistics*, 42(4), 1312–1346.
- Bibinger, M., & Winkelmann, L. (2015). Econometrics of co-jumps in high-frequency data with noise. *Journal of Econometrics*, 184, 361–378.
- Christensen, K., Kinnebrock, S., & Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous. *Journal of Econometrics*, 159, 116–133.
- Cont, R., & Tankov, P. (2004). *Financial modeling with jump process*. London: Chapman & Hall.
- Fissler, T., & Podolskij, M. (2017). Testing the maximal rank of the volatility process for continuous diffusions observed with noise. *Bernoulli*, 23, 3021–3066.
- Fujikoshi, Y., Ulyanov, V., & Shimizu, R. (2010). *Multivariate statistics: High-dimensional and large-sample approximations*. New York: Wiley.
- Gloter, A., & Jacod, J. (2001). Diffusions with measurement errors 1 and 2. *ESAIM: Probability and Statistics*, 5, 225–242.
- Häusler, E., & Luschgy, H. (2015). *Stable convergence and stable limit theorems*. Berlin: Springer.
- Ikedo, N., & Watanabe, S. (1989). *Stochastic differential equations and diffusion processes* (2nd ed.). Amsterdam: North-Holland.
- Jacod, J. (2008). Asymptotic properties of realized power variations and related functionals of semimartingales. *Stochastic Processes and Their Applications*, 118, 517–559.
- Jacod, J., & Podolskij, M. (2013). A test for the rank of the volatility process: The random perturbation approach. *The Annals of Statistics*, 41, 2391–2427.
- Jacod, J., & Protter, P. (2012). *Discretization of processes*. Berlin: Springer.
- Jacod, J., & Todorov, V. (2009). Testing for common arrivals of jumps for discretely observed multidimensional processes. *The Annals of Statistics*, 37, 1792–1838.
- Kitagawa, G. (2010). *Introduction to time series analysis*. Boca Raton: CRC Press.
- Kunitomo, N., & Sato, S. (2013). Separating information maximum likelihood estimation of realized volatility and covariance with micro-market noise. *The North American Journal of Economics and Finance*, 26, 282–309.
- Kunitomo, N. & Sato, S. (2017). “Trend, Seasonality, and Economic Time Series,” A New Approach Using Non-stationary Errors-in-Variables Models. MIMS-RBP Statistics & Data Science Series (SDS-4).
- Kunitomo, N., & Sato, S. (2021). A Robust-filtering method for noisy non-stationary multivariate time series with econometric applications. *Japanese Journal of Statistics and Data Science (JJSD)*. <https://doi.org/10.1007/s42081-020-00102-y>.
- Kunitomo, N., Sato, S., & Kurisu, D. (2018). *Separating information maximum likelihood estimation for high frequency financial data*. Berlin: Springer.
- Kurusu, D. (2018). Power variations and testing for co-jumps: The small noise approach. *Scandinavian Journal of Statistics*, 45(2018), 482–512.
- Li, J., Tauchen, G., & Todorov, V. (2017a). Jump regressions. *Econometrica*, 85, 173–195.
- Li, J., Tauchen, G., & Todorov, V. (2017b). Adaptive estimation of continuous-time regression models using high-frequency data. *Journal of Econometrics*, 200, 36–47.

- Malliavin, P., & Mancino, M. E. (2002). Fourier series method for measurement of multivariate volatilities. *Finance and Stochastics*, 6, 49–61.
- Malliavin, P., & Mancino, M. E. (2009). A fourier transform method for nonparametric estimation of multivariate volatility. *The Annals of Statistics*, 37, 1993–2010.
- Mancino, M. E., Recchioni, M. C., & Sanfelici, S. (2017). *Fourier–Malliavin volatility estimation theory and practice*. Berlin: Springer.
- Pötscher, B. M. (1983). Order estimation in ARMA models by Lagrange multiplier tests. *The Annals of Statistics*, 11, 872–885.
- Robin, J.-M., & Smith, R. J. (2000). Test of rank. *Econometric Theory*, 16, 151–175.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.