Statistica Sinica Preprint No: SS-2022-0259	
Title	On the Estimation of High-Dimensional Integrated Covariance Matrix Based on High-Frequency Data With Multiple Transactions
Manuscript ID	SS-2022-0259
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0259
Complete List of Authors	Moming Wang, Jianhua Hu, Ningning Xia and Yong Zhou
Corresponding Authors	Ningning Xia
E-mails	xia.ningning@mail.shufe.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

On the estimation of high-dimensional integrated covariance matrix based on high-frequency data with multiple transactions

Moming Wang[1], Jianhua Hu[1], Ningning Xia[1], Yong Zhou[2]

[1]*Shanghai University of Finance and Economics and* [2]*East China Normal University*

Abstract: In addition to microstructure noise, the presence of multiple transactions at each recording time is another common feature for high-frequency data. In this paper, we consider the estimation of integrated covariance (ICV) matrix for multiple high-frequency data in a high-dimensional situation where the number of stocks and the "effective" sample size go to infinity proportionally. First, we study the limiting spectral behavior of a pre-averaged version of averaged time-variation adjusted realized covariance (PA-ATVA) matrix based on multiple noisy observations. We show that the PA-ATVA matrix has several desirable properties: it eliminates the effects of microstructure noise and multiple transactions; it allows for rather general dependence structure in the noise process, both cross-sectional and temporal; its LSD depends solely on that of ICV matrix through the Marčenko-Pastur equation. Furthermore, we show that all the aforementioned properties still hold in the presence of asynchronicity. Second, we further propose a nonlinear shrinkage estimator of the ICV matrix based on the PA-ATVA matrix. We show that the proposed estimator is not only asymptotically positive-definite, but also enjoys a desirable estimation efficiency.

At last, simulation and empirical studies demonstrate impressive performance of our proposed estimator.

Key words and phrases: High-dimension, high-frequency, microstructure noise, multiple transactions, random matrix theory.

1. Introduction

1.1 Motivation

The covariance structure of stock market is of great interest to investors and researchers, as it has a critical role in financial problems such as pricing and investment. Suppose that we have p stocks whose log prices at time t are denoted by $\mathbf{X}_t = (X_t^{(1)}, \ldots, X_t^{(p)})^{\mathrm{T}}$, where T denotes the transpose. The following diffusion processes are commonly used to model financial asset price processes:

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\Theta}_t d\mathbf{W}_t, \qquad t \in [0, 1], \tag{1.1}$$

where $(\boldsymbol{\mu}_t) = (\mu_t^{(1)}, \ldots, \mu_t^{(p)})^{\mathrm{T}}$ is a p-dimensional drift process, $(\boldsymbol{\Theta}_t)$ is a $p \times p$ matrix, the so-called covolatility process at time t, and $(\mathbf{W}_t)$ is a p-dimensional standard Brownian motion. The interval $[0, 1]$ represents the time period of interest, such as one trading day. The integrated covariance

(ICV) matrix given by

$$\boldsymbol{\Sigma} := \int_0^1 \boldsymbol{\Theta}_t \boldsymbol{\Theta}_t^{\mathrm{T}} dt \tag{1.2}$$

is of fundamental importance in risk management and portfolio allocation for high-frequency financial data. Due to the unobservable covolatility process $(\boldsymbol{\Theta}_t)$, the estimation of the ICV matrix is an important problem in financial applications, such as portfolio allocation and risk management. However, there are three major issues in estimating the ICV matrix.

This first issue is high dimensionality. Suppose that we can observe the latent log-price processes $(X_{t_i}^{(j)}\}_{j=1}^p$ synchronously at recording time $t_i = i/n$, $i = 0, 1, \ldots, n$, during one trading day. The classical estimator of the ICV matrix is the so-called realized covariance (RCV) matrix, which is defined as

$$\mathrm{RCV} := \sum_{i=1}^{n} \Delta\mathbf{X}_i (\Delta\mathbf{X}_i)^{\mathrm{T}}, \qquad \Delta\mathbf{X}_i = \mathbf{X}_{t_i} - \mathbf{X}_{t_{i-1}}. \tag{1.3}$$

In the case where the dimension p is fixed and the number of observations n goes to infinity, the consistency and the asymptotic normality for the RCV matrix have been well studied, for example, by Andersen and Bollerslev (1998), Andersen et al. (2001), Barndorff-Nielsen and Shephard (2002), and Jacod and Protter (1998). However, in the high-dimensional setting where the dimension p and the observation frequency n go to infinity proportionally, the RCV matrix is no longer consistent. As a result, a large number of studies have worked on the problem of estimating high-dimensional ICV matrix (see Wang and Zou (2010), Fan et al. (2012), Tao et al. (2011), and Zheng and Li (2011), etc). Kim et al. (2018) studied the adaptive thresholding method based on sparse integrated covariance matrix, and Fan et al. (2016), Fan and Kim (2018), Kim et al. (2018) and Kong (2018) studied the covariance matrix estimation based on the approximate factor model.

The second issue is microstructure noise. It is commonly believed that latent asset prices are always contaminated by market microstructure effects, also known as microstructure noise, which is induced by various frictions in the trading process such as asymmetric information among traders and bid-ask spread. The high-frequency accumulation of microstructure noise seriously affects the estimation of the ICV matrix. Consequently, several methods have been developed to deal with microstructure noise, including the two/multi-scale approaches introduced by Aït-Sahalia et al. (2005), Mykland and Zhang (2009), Aït-Sahalia et al. (2010), Aït-Sahalia et al. (2011), Zhang et al. (2005), and Zhang (2006); the realized kernel suggested by Barndorff-Nielsen et al. (2008) and Barndorff-Nielsen et al. (2011); the quasi-maximum likelihood method studied by Xiu (2010); and the pre-averaging approach proposed by Jacod et al. (2009) and Li (2013).

The third issue is multiple transactions. Theoretically, all the abovementioned approaches are only applied to those datasets that contain exactly one transaction during one time interval. The order information is very important since the calculation of RCV matrix is based on the increments $\Delta\mathbf{X}_i$. However, this is not necessarily the case in practice. In high-frequency financial markets, owing to heavy market trading and the limitations of the recording mechanism, multiple transactions often occur at each recording time. Due to the presence of multiple transactions, there may be more than one transactions during one time interval, and the order of consecutive ticks may not be available or even incorrectly recorded. Given the occurrence of multiple transactions, the existing results about ICV matrix should be re-examined. This research is motivated by the fact that few studies have been done on this issue in a high-dimensional noisy environment.

1.2 Contribution

In this paper, we study the estimation of high-dimensional ICV matrix based on high-frequency noisy observations with multiple transactions. As our first contribution, we adopt the pre-averaging approach proposed in Jacod et al. (2009) to deal with the microstructure noise and show that, based on group-averaged observations, the pre-averaged version of averaged time-variation adjusted realized covariance (PA-ATVA) matrix (see (2.8)) has several desirable properties (see Theorem 1): it eliminates the effects of microstructure noise and multiple transactions; it allows for rather general dependence structure in the noise process (both cross-sectional and temporal dependence and even the dependence between the noise and price process); its limiting spectral distribution depends solely on that of the ICV matrix. As our second contribution, in Theorem 2, we show that all the aforementioned desirable properties still hold in the presence of asynchronicity. This is the most exciting result in comparison to previous literatures. As our third contribution, we further propose a nonlinear shrinkage estimator of the ICV matrix based on PA-ATVA matrix. We show that, almost surely, our proposed estimator is positive-definite and enjoys a desirable asymptotic efficiency comparing with the "oracle" one, which relies on the true ICV matrix. All these proofs are under the high-dimensional setting, where the number of stocks and the "effective" sample size grow in the same rate.

The rest of the paper is organized as follows. Section 2 presents the limiting spectral properties of two types of PA-ATVA matrices, one based on noisy observations with multiple transactions and the other based on asynchronicity. In Section 3, a nonlinear shrinkage estimator of the ICV matrix is proposed. Section 4 and 5 demonstrate simulation studies and an analysis of stock market data, respectively. Section 6 contains the conclusions. Some simulations and detailed proofs are provided in the supplementary material.

Notation. We give some notations that will be used throughout this article. Let $\mathbb{C}$, $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ denote the sets of complex, real, integer, and natural numbers, respectively; $\mathbb{C}^+$, a subset of $\mathbb{C}$, contains positive imaginary parts. All vectors are column vectors, and we use $|\cdot|$ to denote the Euclidean norm for vectors. The transpose of any matrix $\mathbf{A}$ is denoted by $\mathbf{A}^{\mathrm{T}}$. From random matrix theory, the empirical spectral distribution (ESD) is defined as

$$F^{\mathbf{A}}(x) = \frac{1}{p}\sum_{j=1}^{p} \mathbf{I}(\lambda_j(\mathbf{A}) \leq x), \quad \text{for} \quad x \in \mathbb{R}, \tag{1.4}$$

where $\mathbf{I}(\cdot)$ is an indicator function and $\lambda_j(\mathbf{A})$ denotes the jth largest eigenvalue of a Hermitian matrix $\mathbf{A}$. The limit of ESD is referred to as the limiting spectral distribution (LSD), if it exists. For any distribution function F, let $m_F(\cdot)$ denote its Stieltjes transform defined as

$$m_F(z) = \int \frac{1}{\lambda - z} dF(\lambda), \qquad \text{for } z \in \mathbb{C}^+ := \{z \in \mathbb{C} : \Im(z) > 0\},$$

where $\Im(z)$ is the imaginary part of z. We use notation $\lfloor \cdot \rfloor$ to indicate rounding down to the nearest integer.

2. The limiting spectral distributions of PA-ATVA matrices

2.1 The setup

The following model settings are assumed in the case of high-frequency data with multiple transactions. Suppose that the latent log price process $(\mathbf{X}_t)$ follows model (1.1) and the observed contaminated process $(\mathbf{Y}_t)$ is assumed to follow the additive model

$$\mathbf{Y}_t = \mathbf{X}_t + \boldsymbol{\varepsilon}_t, \qquad \text{for } t \in [0,1], \tag{2.5}$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_t^{(1)}, \ldots, \varepsilon_t^{(p)})^{\mathrm{T}}$ denotes the noise process. For the sake of convenience, we will refer to any process in this work as $(\mathbf{V}_t)$, regardless of whether it is the latent process $(\mathbf{X}_t)$, the contaminated process $(\mathbf{Y}_t)$, or the noise process $(\boldsymbol{\varepsilon}_t)$. In a given time interval $[0,1]$, one trading day, for any process $(\mathbf{V}_t)$, suppose that $(\mathbf{V}_t)$ can be observed at time points $t_i = i/n$ and there are $L_i (L_i \geq 1)$ transactions during each time interval $(t_{i-1}, t_i]$, for $i = 1, \ldots, n$. Let $V_{T_{i-1}+j}$ be the jth observation at transaction time $s_{T_{i-1}+j}$ during time interval $(t_{i-1}, t_i]$ with $T_0 \equiv 0$ and $T_i = \sum_{k=1}^{i} L_k$, for $j = 1, 2, \ldots, L_i$ and $i = 1, 2, \ldots, n$. That is, L_i transactions at time points $\{s_{T_{i-1}+j}, j = 1, 2, \ldots, L_i\}$ occurred during $(t_{i-1}, t_i]$, where $t_{i-1} < s_{T_{i-1}+1} < s_{T_{i-1}+2} < \cdots < s_{T_{i-1}+L_i} = s_{T_i} \leq t_i$. The observations at each recording time point $\{t_i\}$ are as follows:

at time t_0: $\mathbf{V}_{t_0} = \mathbf{V}_0$;

at time t_1: $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_{T_1}$;

$\vdots$

at time t_n: $\mathbf{V}_{T_{n-1}+1}, \mathbf{V}_{T_{n-1}+2}, \ldots, \mathbf{V}_{T_n}$.

Theoretically, these observations occur consecutively during time interval $(t_{i-1}, t_i]$. However, in practice, the observations at each recording time t_i lose their order of arrival owing to the recording mechanism. The L_i observations occurring during time interval $(t_{i-1}, t_i]$ are only recorded at time t_i. Specifically,

$$\mathbf{V}_0, \quad \underbrace{\mathbf{V}_1, \ldots, \mathbf{V}_{T_1}}_{L_1 \text{ observations at } t_1}, \ldots, \underbrace{\mathbf{V}_{T_{i-1}+1}, \ldots, \mathbf{V}_{T_i}}_{L_i \text{ observations at } t_i}, \ldots, \underbrace{\mathbf{V}_{T_{n-1}+1}, \ldots, \mathbf{V}_{T_n}}_{L_n \text{ observations at } t_n}.$$

One commonly used method to deal with multiple transactions involves taking averages of the multiple transaction prices at each time point t_i, that is,

$$\overline{\mathbf{V}}_i := \frac{1}{L_i} \sum_{j=1}^{L_i} \mathbf{V}_{T_{i-1}+j}, \qquad i = 1, \ldots, n, \tag{2.6}$$

and using $\overline{\mathbf{V}}_i$ to approximate the observation at time point t_i. Moreover, to deal with the microstructure noise, we adopt the pre-averaging approach with a window length h and further introduce the following notations,

$$\widetilde{\overline{\mathbf{V}}}_i := \frac{1}{h} \sum_{j=1}^{h} \overline{\mathbf{V}}_{(i-1)h+j} \quad \text{and} \quad \Delta \widetilde{\overline{\mathbf{V}}}_{2i} := \widetilde{\overline{\mathbf{V}}}_{2i} - \widetilde{\overline{\mathbf{V}}}_{2i-1}.$$

In this way, the observed return based on the pre-averaged price becomes

$$\Delta\widetilde{\widetilde{\mathbf{Y}}}_{2i} = \Delta\widetilde{\widetilde{\mathbf{X}}}_{2i} + \Delta\widetilde{\widetilde{\boldsymbol{\varepsilon}}}_{2i}.$$

2.2 The LSD of PA-ATVA matrix

Definition 1. Suppose that $(\mathbf{X}_t)$ is a p-dimensional process satisfying (1.1) and $\mathbf{\Theta}_t$ is càdlàg. Then $(\mathbf{X}_t)$ belongs to class $\mathcal{C}$ if, almost surely, there exist $\gamma_t \in D([0,1];\mathbb{R})$ and $\mathbf{\Lambda}$ a $p \times p$ matrix satisfying $\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}) = p$ such that

$$\mathbf{\Theta}_t = \gamma_t \mathbf{\Lambda}, \tag{2.7}$$

where $D([0,1];\mathbb{R})$ is the space of càdlàg functions from $[0,1]$ to $\mathbb{R}$.

For class $\mathcal{C}$ processes, Zheng and Li (2011) proposes the time-variation adjusted realized covariance (TVA) matrix, which is given by

$$\mathrm{TVA} := \frac{\mathrm{tr(RCV)}}{n} \sum_{i=1}^{n} \frac{\Delta\mathbf{X}_i(\Delta\mathbf{X}_i)^{\mathrm{T}}}{|\Delta\mathbf{X}_i|^2}.$$

The advantage of considering TVA matrix instead of RCV matrix is that the LSD of TVA depends solely on that of ICV through Marčenko-Pastur equation. Hence, the ESD of ICV can be recovered by using existing methods (such as Karoui (2008) etc).

In this paper, to investigate the effect of multiple transactions based on noisy high-frequency data, we start by studying the pre-averaged version of TVA matrix based on the group-averaged observations $(\overline{\mathbf{Y}}_i)_{i=1}^n$, which is defined as follows. Let window length $h = \lfloor \xi n^\beta \rfloor$ with $\xi \in (0, \infty)$ and $\beta \in (1/2, 1)$. Take the "effective" sample size $M = \lfloor n/(2h) \rfloor$. The pre-averaged version of averaged TVA (PA-ATVA) matrix is then defined as

$$\mathcal{B}_M := 3\frac{\sum_{i=1}^M |\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}|^2}{M} \cdot \sum_{i=1}^M \frac{\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}(\Delta\widetilde{\overline{\mathbf{Y}}}_{2i})^{\mathrm{T}}}{|\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}|^2} = 3\frac{\sum_{i=1}^M |\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}|^2}{p} \cdot \widetilde{\boldsymbol{\Xi}}, \qquad (2.8)$$

where

$$\widetilde{\boldsymbol{\Xi}} := \frac{p}{M} \sum_{i=1}^M \frac{\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}(\Delta\widetilde{\overline{\mathbf{Y}}}_{2i})^{\mathrm{T}}}{|\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}|^2}. \qquad (2.9)$$

One key observation is that the window length h has a higher order than $\sqrt{n}$, which enables us to asymptotically eliminate the effect of microstructure noise. To study the behavior of $\mathcal{B}_M$ based on noisy and multiple observations, we require some assumptions regarding the noise process. Recall that the definition of ρ-mixing coefficients is as follows.

Definition 2. For a stationary time series $(U_k), k \in \mathbb{Z}$, let $\mathcal{F}_j^\ell$ be the σ-field generated by the random variables $(U_k : -\infty \le j \le k \le \ell \le \infty)$. The ρ-mixing coefficients are defined as

$$\rho(r) = \sup_{f \in \mathcal{L}^2(\mathcal{F}_{-\infty}^0),\ \ g \in \mathcal{L}^2(\mathcal{F}_r^\infty)} |\mathrm{Corr}(f, g)|, \quad \text{for} \quad r \in \mathbb{N},$$

where, for any probability space Ω, $\mathcal{L}^2(\Omega)$ refers to the space of square-integrable, Ω-measurable random variables.

We now state our assumptions.

(A.i) For all p, $(\mathbf{X}_t)$ is a p-dimensional process in class $\mathcal{C}$ for some drift process $\boldsymbol{\mu}_t = (\mu_t^{(1)}, \ldots, \mu_t^{(p)})^{\mathrm{T}}$ and covolatility process $(\boldsymbol{\Theta}_t) = (\gamma_t \boldsymbol{\Lambda})$;

(A.ii) there exists a $C_0 < \infty$ such that for all p and all $q = 1, \ldots, p$, $|\mu_t^{(q)}| \leq C_0$ for all $t \in [0, 1)$ almost surely;

(A.iii) there exists a $0 \leq \delta_1 < 1/2$ and a sequence of index sets $\mathcal{I}_p$ satisfying $\mathcal{I}_p \subset \{1, \ldots, p\}$ and $\#\mathcal{I}_p = O(p^{\delta_1})$ such that (γ_t) may depend on $\mathbf{W}_t$ but only on $(W_t^{(q)} : q \in \mathcal{I}_p)$; moreover, there exists a $C_1 < \infty$ such that for all p, $|\gamma_t| \in (1/C_1, C_1)$ for all $t \in [0, 1)$, almost surely;

(A.iv) there exists $C_2 < \infty$ such that for all p and all l, the individual volatilities $\sigma_t^{(l)} = \sqrt{(\gamma_t)^2 \sum_{k=1}^{p} (\Lambda_{lk})^2} \in (1/C_2, C_2)$ for all $t \in [0, 1]$ almost surely; in addition, (γ_t) converges uniformly to a nonzero process (γ_t^*) that is piecewise continuous with finitely many jumps almost surely;

(A.v) there exists $C_3 < \infty$ and $0 \leq \delta_2 < 1/2$ such that for all p, $\|\mathrm{ICV}\| \leq C_3 p^{\delta_2}$ almost surely;

(A.vi) the δ_1 in (A.iii) and δ_2 in (A.v) satisfy that $\delta_1 + \delta_2 < 1/2$;

(A.vii) almost surely, as $p \to \infty$, the ESD of $\breve{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ converges in distribution to a probability distribution $\breve{H}$;

(A.viii) For all $q = 1, \ldots, p$, the noise $(\varepsilon_t^{(q)})$ is a stationary time series with mean 0 and has bounded 4νth moments and ρ-mixing coefficients $\rho^{(q)}(r)$ satisfying $\max_{q=1,\ldots,p} \rho^{(q)}(r) = O(r^{-\nu})$ for some integer $\nu > 2$;

(A.ix) $h = \lfloor \xi n^{\beta} \rfloor$ for some $\xi \in (0, \infty)$ and $\beta \in ((3+\nu)/(2\nu+2), 1)$, and $M = \lfloor n/(2h) \rfloor$ satisfy that $\lim_{p\to\infty} p/M = c > 0$, where ν is the integer in Assumption (A.viii);

(A.x) $\max_{1\le i\le n} n(s_{T_i} - t_i) \to 0$ almost surely as $p \to \infty$;

(A.xi) there exists a constant $L^* < \infty$ such that $1 \le \inf L_i \le \sup L_i \le L^*$, almost surely;

Assumptions (A.i)–(A.vi) are reasonable for the underlying log-price process proposed in Zheng and Li (2011). In such a case, no sparsity assumption on the ICV matrix is needed, the dependence between the covolatility process and the Brownian motion in Assumption (A.iii) allows the leverage effect to be captured, and the jumps are considered in Assumption (A.iv). Assumptions (A.vii) is a standard for use in high-dimensional settings in random matrix theory. Assumption (A.viii) is a quite "mild" assumption as pointed out by Xia and Zheng (2018). It allows for not only dependence within the noise process, both cross-sectional and temporal, but also dependence between the noise and price processes. Assumption (A.ix) follows from assumption (D.vii) in Xia and Zheng (2018), which is a standard condition in random matrix theory. Assumption (A.x) assumes that the latest multiple-transaction time in each time interval is close enough to the recording time, which is a reasonable condition for high-frequency data, especially facing with ultra-high-frequency data. Assumption (A.xi) considers the number of multiple transactions, which are assumed to be finite. Then, we have the following convergence result for the PA-ATVA matrix $\mathcal{B}_M$.

Theorem 1. *Suppose that Assumptions* (A.i)–(A.xi) *hold. Then, as* $p \to \infty$*, the ESDs of ICV and* $\mathcal{B}_M$ *converge almost surely to probability distributions* H *and* $F^{\mathcal{B}}$*, respectively, where*

$$H(x) = \breve{H}(x/\theta), \quad \text{for all } x \geq 0 \text{ with } \theta = \int_0^1 (\gamma_t^*)^2 dt, \tag{2.10}$$

and $F^{\mathcal{B}}$ *is determined by* H *in that its Stieltjes transform,* $m_{\mathcal{B}}(z)$*, satisfies the Marčenko-Pastur equation,*

$$m_{\mathcal{B}}(z) = \int_{\tau \in \mathbb{R}} \frac{1}{\tau(1 - c(1 + z m_{\mathcal{B}}(z))) - z} dH(\tau), \quad \text{for } z \in \mathbb{C}^+. \tag{2.11}$$

Compared with Theorem 2 in Zheng and Li (2011) and Theorem 2.3 in Xia and Zheng (2018), our Theorem 1 shows that the PA-ATVA matrix $\mathcal{B}_M$

eliminates not only the effect of microstrucute noise but also the effect of multiple transactions. The LSDs of ICV and $\mathcal{B}_M$ are solely related through Marčenko-Pastur equation.

2.3 Asynchronous trading

In this subsection, we consider a setting where the observations are asynchronous. In addition to microstructure noise and multiple transactions, asynchronous trading is another challenge for high-dimensional and high-frequency data analysis. In practice, different stocks usually have different numbers of transactions during one time interval. Take an example of Apple, Cisco Systems, and Microsoft; on November 4, 2016, there were 491, 166, and 217 transactions in the first trading seconds, respectively. Denote $L_i^{(q)}$ the number of multiple transactions for stock q, $1 \leq q \leq p$, during recording interval $(t_{i-1}, t_i]$. For any process $(\mathbf{V}_t)$, let $V_{i,j}^{(q)}$ be the observation of the jth transaction for stock q during time interval $(t_{i-1}, t_i]$ with $t_i = i/n$. A figure in the supplementary material shows a simplified version of true transactions and recording mechanism under asynchronous trading.

Let $T_i^{(q)} = \sum_{k=1}^{i} L_k^{(q)}$ for $q = 1, \ldots, p$ and $i = 1, \ldots, n$. The true transaction time of $V_{i,j}^{(q)}$ is denoted as $s^{(q)}_{T^{(q)}_{i-1}+j}$, for $j = 1, \ldots, L_i^{(q)}$, satisfying $t_{i-1} < s^{(q)}_{T^{(q)}_{i-1}+1} < \cdots < s^{(q)}_{T^{(q)}_{i-1}+L^{(q)}_i} = s^{(q)}_{T^{(q)}_i} \leq t_i$. With asynchronous trading, the average of multiple observations at each recording time t_i is denoted as

$$\overline{\mathbf{V}}_i^* = \left(\sum_{j=1}^{L_i^{(1)}} \frac{1}{L_i^{(1)}} V_{i,j}^{(1)}, \ldots, \sum_{j=1}^{L_i^{(p)}} \frac{1}{L_i^{(p)}} V_{i,j}^{(p)} \right)^{\mathrm{T}},$$

and the pre-averaging observation becomes

$$\widetilde{\overline{\mathbf{V}}}_i^* = \frac{1}{h} \sum_{j=1}^{h} \overline{\mathbf{V}}_{(i-1)h+j}^*,$$

for $i = 1, \ldots, n$. Define the increments $\Delta\widetilde{\overline{\mathbf{V}}}_{2i}^* = \widetilde{\overline{\mathbf{V}}}_{2i}^* - \widetilde{\overline{\mathbf{V}}}_{2i-1}^*$. Using the above notation, the PA-ATVA matrix is then rewritten as

$$\mathcal{B}_M^* = 3\frac{\sum_{i=1}^{M} |\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}^*|^2}{M} \cdot \sum_{i=1}^{M} \frac{\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}^* (\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}^*)^{\mathrm{T}}}{|\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}^*|^2}. \tag{2.12}$$

We further assume Assumptions (B.i) and (B.ii) as follows:

(B.i) there exists a constant $L^* < \infty$ such that $1 \leq \inf_{i,q} L_i^{(q)} \leq \sup_{i,q} L_i^{(q)} \leq L^*$, almost surely;

(B.ii) $\max_{1\leq i\leq n, 1\leq q\leq p} nh(s_{T_i^{(q)}}^{(q)} - t_i) \to 0$, almost surely, as $p \to \infty$.

Assumption (B.i) is parallel to Assumption (A.xi). Assumption (B.ii) requires a faster convergence rate than Assumption (A.x), yet Assumption (B.ii) is still feasible when dealing with ultra-high-frequency dataset. With tick-by-tick transaction data, one recording time interval can be as little as one millisecond. Because the distances between them are too small to capture, we can infer that $s_{T_i^{(q)}}^{(q)} \approx t_i$ in practice.

Theorem 2. *Suppose that Assumptions* (A.i)–(A.ix) *and* (B.i)–(B.ii) *hold. Then, as* $p \to \infty$*, the ESDs of ICV and* $\mathcal{B}_M^*$ *converge almost surely to probability distributions* H *and* $F^{\mathcal{B}^*}$*, respectively, where* H *satisfies* (2.10) *and* $F^{\mathcal{B}^*}$ *is determined by* H *in that its Stieltjes transform* $m_{\mathcal{B}^*}(z)$ *satisfies the Marčenko-Pastur equation,*

$$m_{\mathcal{B}^*}(z) = \int_{\tau \in \mathbb{R}} \frac{1}{\tau(1 - c(1 + z m_{\mathcal{B}^*}(z))) - z} dH(\tau), \quad \text{for } z \in \mathbb{C}^+.$$

Based on Theorem 1, Theorem 2 further states that the (rescaled) PA-ATVA matrix $\mathcal{B}_M^*$ is not influenced by the effect of asynchronous trading, which is an important property comparing with existing literatures. Hence, the matrix $\mathcal{B}_M^*$ provides a robust basis for making further inference of the ICV matrix.

3. The proposed estimator

We can see from Theorem 2 that the LSD of the ICV matrix is not the same as the LSD of matrix $\mathcal{B}_M^*$ in general. Furthermore, Theorem 2 indicates that the extreme eigenvalues of matrix $\mathcal{B}_M^*$ are more extreme than their ICV counterparts, following a similar discussion as in Bai and Yin (1993); Yin et al. (1998). That is, the smallest eigenvalue of matrix $\mathcal{B}_M^*$ is smaller than the smallest eigenvalue of the ICV matrix, while the maximum eigenvalue is larger than the largest eigenvalue of the ICV matrix. As evidenced by these findings, matrix $\mathcal{B}_M^*$ is still not a good estimator of the target ICV matrix. This leads to our next goal, which is to develop a suitable ICV estimator based on noisy high-frequency data in the presence of multiple-transactions.

Ledoit and Wolf (2012) offers a class of rotation-equivariant estimators that keep the sample covariance matrix's eigenvectors but nonlinearly shrink its eigenvalues without making any assumptions about the population covariance matrix's structure. Furthermore, the authors published a series of publications based on i.i.d. samples to build an estimator of the ideal nonlinear shrinkage function. Realizing the difficulties of estimating the unknown parameters in the optimal nonlinear shrinkage estimator in the high-frequency framework, in this section, we adopt the approach of Abadir et al. (2014); Lam (2016) to examine the nonparametric eigenvalue-regularized covariance estimator by partitioning the dataset into two parts.

3.1 The estimator

Suppose that we have tick-by-tick high-frequency observations $(\{\mathbf{Y}_{i,j}^{(q)}\}_{q=1}^p)$ as assumed in Section 2.3. Denote the matrix of pre-averaging returns $\Delta\boldsymbol{\Upsilon} = \{\Delta\widetilde{\overline{\mathbf{Y}}}_2^*, \ldots, \Delta\widetilde{\overline{\mathbf{Y}}}_{2M}^*\}$. We split the data $\Delta\boldsymbol{\Upsilon}$ into two parts $\{\Delta\boldsymbol{\Upsilon}_1, \Delta\boldsymbol{\Upsilon}_2\}$ with $\Delta\boldsymbol{\Upsilon}_j$ having size $p \times M_j$ for $j = 1, 2$, such that $M_1 + M_2 = M$. Define

$$\widetilde{\boldsymbol{\Xi}}_j^* := \frac{p}{M_j} \sum_{i \in J_j} \frac{\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}^* (\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}^*)^{\mathrm{T}}}{|\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}^*|^2},$$

where $J_j = \{i : \Delta\widetilde{\overline{\mathbf{Y}}}^*_{2i} \in \Delta\boldsymbol{\Upsilon}_j\}$ for $j = 1, 2$ and write the spectral decomposition of $\widetilde{\boldsymbol{\Xi}}^*_1$ as $\widetilde{\boldsymbol{\Xi}}^*_1 = \mathbf{U}_1\widetilde{\boldsymbol{\Lambda}}_1\mathbf{U}_1^{\mathrm{T}}$, where $\widetilde{\boldsymbol{\Lambda}}_1$ is a diagonal matrix of eigenvalues of $\widetilde{\boldsymbol{\Xi}}_1$ and $\mathbf{U}_1 = (\mathbf{u}_{11}, \mathbf{u}_{12}, \ldots, \mathbf{u}_{1p})$ is the eigenmatrix with each column $\mathbf{u}_{1i}$ being the corresponding eigenvector. Then we define our estimator as

$$\widehat{\boldsymbol{\Sigma}} := 3\frac{\sum_{i=1}^{M} |\Delta\widetilde{\overline{\mathbf{Y}}}^*_{2i}|^2}{p}\widehat{\boldsymbol{\Xi}}^*, \quad \widehat{\boldsymbol{\Xi}}^* = \mathbf{U}_1\mathrm{diag}\left(\mathbf{U}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Xi}}^*_2\mathbf{U}_1\right)\mathbf{U}_1^{\mathrm{T}}, \tag{3.13}$$

where $\mathrm{diag}(\mathbf{A})$ sets all non-diagonal elements of matrix $\mathbf{A}$ to zero. Because there is no previous information on the orientation of the targeting ICV matrix, the eigenvectors of $\widetilde{\boldsymbol{\Xi}}^*_1$ are kept, and the eigenvalues are shrunk by splitting the dataset into two parts.

3.2 Estimation efficiency

We introduce the following "oracle" estimator to develop the estimation efficiency of our estimator

$$\boldsymbol{\Sigma}_{\mathrm{oracle}} := \mathbf{U}\mathrm{diag}(\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U})\mathbf{U}^{\mathrm{T}},$$

where $\boldsymbol{\Sigma}$ is the ICV matrix given in (1.2) and $\mathbf{U}$ is the eigenmatrix of

$$\widetilde{\boldsymbol{\Xi}}^* = \frac{p}{M} \cdot \sum_{i=1}^{M} \frac{\Delta\widetilde{\overline{\mathbf{Y}}}^*_{2i}(\Delta\widetilde{\overline{\mathbf{Y}}}^*_{2i})^{\mathrm{T}}}{|\Delta\widetilde{\overline{\mathbf{Y}}}^*_{2i}|^2} \tag{3.14}$$

such that the spectral decomposition of $\widetilde{\boldsymbol{\Xi}}^*$ is $\mathbf{U}\widetilde{\boldsymbol{\Lambda}}\mathbf{U}^{\mathrm{T}}$. The matrix $\boldsymbol{\Sigma}_{\mathrm{oracle}}$ is called "oracle" because on one hand it depends on the unknown ICV matrix, on the other hand it uses the full set of data for obtaining the eigenmatrix $\mathbf{U}$ and solves the following optimization problem

$$\min_{\widetilde{\mathbf{D}} \text{ is diagonal}} \|\mathbf{U}\widetilde{\mathbf{D}}\mathbf{U}^{\mathrm{T}} - \mathbf{\Sigma}\|_F, \tag{3.15}$$

where $\widetilde{\mathbf{D}}$ is a diagonal matrix and $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^{\mathrm{T}})$ is the Frobenius norm of a matrix $\mathbf{A}$. With this "oracle" estimator, we consider the efficiency loss of our estimator $\widehat{\mathbf{\Sigma}}$ as

$$EL(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) = 1 - \|\mathbf{\Sigma} - \mathbf{\Sigma}_{\mathrm{oracle}}\|_F^2 / \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_F^2.$$

When compared to the "oracle" estimator $\mathbf{\Sigma}_{\mathrm{oracle}}$, $EL(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) \leq 0$ indicates that the estimator $\widehat{\mathbf{\Sigma}}$ causes a smaller Frobenius loss or is doing as well as $\mathbf{\Sigma}_{\mathrm{oracle}}$, and vice versa.

Next, to investigate the limiting properties of the proposed estimator $\widehat{\mathbf{\Sigma}}$, we assume the following assumptions.

(C.i) All the eigenvalues of $\breve{\mathbf{\Sigma}} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$ are bounded uniformly from 0 and infinity for $t \in [0, 1]$.

(C.ii) The drift process satisfy that $p^{-1/2}|\boldsymbol{\mu}_t| \to 0$ for all $t \in [0, 1]$ almost surely.

(C.iii) The covolatility process (γ_t) is independent of $(\mathbf{W}_t)$.

(C.iv) The number M_2 satisfy that $\sum_{p \geq 1} p M_2^{-5} < \infty$.

(C.v) Denote $h = \lfloor \xi n^{\beta} \rfloor$ for some $\xi \in (0, \infty)$ and $\beta \in (\frac{2\nu+2}{3\nu+2}, 1)$, and

$M = \lfloor n(2h)^{-1} \rfloor$ satisfying that $\lim_{p\to\infty} p/M = \varrho > 0$ where the integer ν is the integer given in Assumption (A.viii).

To establish second-order consistency, we need stronger conditions in Assumptions (C.i)–(C.iii) than in Assumptions (A.ii)–(A.v), but they are still far weaker than the assumptions in Lam et al. (2017). All of the assumptions are merely necessary for the theoretical findings to be obtained. We will demonstrate via simulation studies that the proposed estimator's finite sample performance is outstanding even when the assumptions are violated. Assumption (C.iv) assumes the split location number, in which we suppose that M_2, the size of second dataset, goes to infinity with diverging rate at least faster than $O(p^{(2+\alpha)/5})$ for some $\alpha > 0$. The Assumption (C.v) is a standard requirement in random matrix theory.

Theorem 3. *Under Assumptions in Theorem 2, further we suppose that Assumptions* (C.i)–(C.v) *hold. Then, we have almost surely, the proposed estimator* $\widehat{\mathbf{\Sigma}}$ *is asymptotically positive-definite.*

Corollary 1. *Suppose that the assumptions in Theorem 3 hold. Then*

$$\max_{1\le\ell\le p} \left| diag(\mathbf{U}_1^{\mathrm{T}}\widetilde{\mathbf{\Xi}}_2^{*}\mathbf{U}_1)\, diag^{-1}(\mathbf{U}_1^{\mathrm{T}}\check{\mathbf{\Sigma}}\mathbf{U}_1) - 1 \right| \;\to\; 0, \qquad \textit{almost surely.}$$

Remark 1. Theorem 3 shows that the proposed estimator is always asymptotically positive-definite even when the dimension p is larger than the effective sample size M. It is critical in portfolio allocation since some portfolio strategy (see the GMP strategy in empirical study) may require an invertible estimator for ICV matrix which is not always satisfied for example the RCV or PA-ATVA matrix.

Remark 2. The factor model is a common approach in modeling high-frequency data. See Fan and Kim (2018) for example. To account for the effect of common factors, we consider the following factor model

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \xi_t \boldsymbol{\Lambda}_f d\mathbf{W}_t^* + \gamma_t \boldsymbol{\Lambda}_B d\mathbf{W}_t,$$

where $(\boldsymbol{\mu}_t)$ is a $p \times 1$ drift process, (ξ_t) and (γ_t) are processes defined on $D([0,1];\mathbb{R})$, $\boldsymbol{\Lambda}_f$ is a $p \times r$ factor loading matrix, $\boldsymbol{\Lambda}_B$ is a $p \times p$ matrix satisfying $\mathrm{tr}(\boldsymbol{\Lambda}_B \boldsymbol{\Lambda}_B^{\mathrm{T}}) = p$, and $\mathbf{W}_t^*$ and $\mathbf{W}_t$ are r-dimensional and p-dimensional independent Brownian motions, respectively. The ICV matrix is then given by $\boldsymbol{\Sigma} = \int_0^1 \xi_t^2 dt \cdot \boldsymbol{\Lambda}_f \boldsymbol{\Lambda}_f^{\mathrm{T}} + \int_0^1 \gamma_t^2 dt \cdot \boldsymbol{\Lambda}_B \boldsymbol{\Lambda}_B^{\mathrm{T}}$. The matrix $\boldsymbol{\Sigma}_f = \int_0^1 \xi_t^2 dt \cdot \boldsymbol{\Lambda}_f \boldsymbol{\Lambda}_f^{\mathrm{T}}$ (with small rank $r < \infty$) accounts for the influence of common factors on the ICV matrix. (1) For a specific setting, when the processes (ξ_t) and (γ_t) are identical, that is the matrix $\boldsymbol{\Lambda}$ in (2.7) admits the decomposition $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_f, \boldsymbol{\Lambda}_B)$, under the condition of $p^{-1}\mathrm{tr}(\boldsymbol{\Lambda}_f \boldsymbol{\Lambda}_f^{\mathrm{T}}) \to \beta^* \in [0, \infty)$, one can show that Theorem 3 still holds following a similar discussion as in Lemma 1 of Lam (2016) (see the supplementary material). (2) For a general setting, the desirable properties of our proposed estimator may not be satisfied. Hence, the estimation of ICV matrix for high-frequency factor models needs to be further investigated.

Theorem 4. *Suppose that all assumptions in Theorem 3 hold. Additionally, if $M_1/M \to 1$ and $M_2 \to \infty$, then we have $EL(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}) \le 0$ almost surely, provided that $p^{-1}\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{oracle}\|_F^2 \not\to 0$ almost surely.*

The condition $p^{-1}\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\mathrm{oracle}}\|_F^2 \not\to 0$ implies that the case $\breve{\boldsymbol{\Sigma}} = \mathbf{I}_p$ is excluded, which is a plausible assumption. In the financial markets, the correlation of stocks is a well-known fact. The matrix $\boldsymbol{\Sigma}_{\mathrm{oracle}}$ is an optimal estimator by solving the optimization problem (3.15) in the class $\mathcal{S} = \{\mathbf{U}\widetilde{\mathbf{D}}\mathbf{U}^{\mathrm{T}},$ where $\widetilde{\mathbf{D}}$ is diagonal and $\mathbf{U}$ is the eigenmatrix of $\widetilde{\boldsymbol{\Xi}}^*$ in (3.14)$\}$. There may be an estimator $\widehat{\boldsymbol{\Sigma}}$ with smaller efficiency loss when $\widehat{\boldsymbol{\Sigma}} \notin \mathcal{S}$. Theorem 4 indicates that our proposed estimator $\widehat{\boldsymbol{\Sigma}}$ works better or at least performs as well as $\boldsymbol{\Sigma}_{\mathrm{oracle}}$ in terms of Frobenius loss. That is why our estimator works well in a variety of situations, including spiked and factor models. Simulation studies support our approach.

3.3 Improvement with averaging

To improve the performance of $\widehat{\boldsymbol{\Sigma}}$ in (3.13), we can permute the return data during the time of interest B times. At the kth permutation, we split the increments of $\Delta\boldsymbol{\Upsilon}$ into two parts $\{\Delta\boldsymbol{\Upsilon}_1^{(k)}, \Delta\boldsymbol{\Upsilon}_2^{(k)})\}(k = 1, \ldots, B)$ with

$\Delta\boldsymbol{\Upsilon}_j^{(k)}$ having size $p \times M_j$ for $j = 1, 2$, such that $M_1 + M_2 = M$. Let $(\widetilde{\boldsymbol{\Xi}}_j^*)^{(k)}$ denotes $\widetilde{\boldsymbol{\Xi}}_j^*$ at kth permutation with spectral decomposition $\mathbf{U}_j^{(k)}\widetilde{\boldsymbol{\Lambda}}_j^{(k)}(\mathbf{U}_j^{(k)})^{\mathrm{T}}$ for $k = 1, \ldots, B, j = 1, 2$. Then the averaged estimator of ICV matrix is defined as

$$\widehat{\boldsymbol{\Sigma}}_B := 3\frac{\sum_{i=1}^{M} |\Delta\widetilde{\overline{\mathbf{Y}}}_{2i}^*|^2}{p} \cdot \frac{1}{B}\sum_{k=1}^{B}(\widehat{\boldsymbol{\Xi}}^*)^{(k)}, \tag{3.16}$$

where $(\widehat{\boldsymbol{\Xi}}^*)^{(k)} = \mathbf{U}_1^{(k)}\mathrm{diag}\left((\mathbf{U}_1^{(k)})^{\mathrm{T}}(\widetilde{\boldsymbol{\Xi}}_2^*)^{(k)}\mathbf{U}_1^{(k)}\right)(\mathbf{U}_1^{(k)})^{\mathrm{T}}$. In practice, following the criterion of split location proposed by Lam (2016), we simply set $B = 50$ and M_1 is given by minimizaing the following problem

$$\arg\min_{M_1 \in \mathcal{G}} \left|\left|\frac{1}{B}\sum_{k=1}^{B}\{(\widehat{\boldsymbol{\Xi}}^*)^{(k)} - (\widetilde{\boldsymbol{\Xi}}_2^*)^{(k)}\}\right|\right|_F^2, \tag{3.17}$$

where

$$\mathcal{G} = \left[2M^{1/2}, 0.2M, 0.4M, 0.6M, 0.8M, M - 2.5M^{1/2}, M - 1.5M^{1/2}\right]$$

is the candidate set for finite sample performance.

Theorem 5. *Suppose that all assumptions in Theorem 4 hold. For any finite number B, we have that the averaged estimator of ICV matrix defined in* (3.16) *is asymptotically positive-definite almost surely. Moreover, we have $EL(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_B) \leq 0$ almost surely.*

Theorem 5 states that the permutation has no effect on the proposed estimator's desirable features. In general, averaging is a good way to get a more reliable estimator. As a result, in empirical studies, we employ the averaged version of estimation.

4. Simulation Studies

In this section, we conduct some simulation studies to assess the performance of the proposed estimator (3.16) in a variety of scenarios, including spiked models (the largest few eigenvalues of ICV matrix differing from the remainder) and factor models. Firstly, we assume that the latent log-price process $(\mathbf{X}_t)$ follows $d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \gamma_t \boldsymbol{\Lambda} d\mathbf{W}_t$, where $\boldsymbol{\mu}_t$ are i.i.d. $N(0.02, (0.002)^2\mathbf{I}_p)$. We introduce two kinds of (γ_t). The first one is a piecewise constant function with

$$\gamma_t = \begin{cases} 0.01 \times 7^{1/2}, & 0 \le t < 1/4 \text{ or } 3/4 \le t \le 1 \\ 0.01, & 1/4 \le t \le 3/4. \end{cases} \tag{4.18}$$

The second one is the U-shaped stochastic process (γ_t) defined as follows:

$$d\gamma_t = -\rho(\gamma_t - \mu_t)dt + \sigma d\widetilde{W}_t, \quad \text{for} \quad t \in [0, 1], \tag{4.19}$$

where $\rho = 10, \sigma = 0.05, \mu_t = \sqrt{0.0009 + 0.0008\cos(2\pi t)}$, and the process $\widetilde{W}_t = \sum_{i=1}^p W_t^{(i)}/\sqrt{p}$ with $W_t^{(i)}$ being the ith component of the Brownian motion $(\mathbf{W}_t)$ that drives the price process. Note that (γ_t) dependents on all components of the Brownian motion and Assumption (A.iii) and Assumption (C.iii) are violated, but the proposed estimator $\widehat{\boldsymbol{\Sigma}}_B$ still works quite well.

- Case I: assume that γ_t is a process of form (4.19), $\boldsymbol{\Lambda} = (0.5^{|i-j|})_{i,j=1,\ldots,p}$;
- Case II: assume that γ_t is a U-shaped process of form (4.19) and set $\boldsymbol{\Lambda} = (0.5^{|i-j|})_{i,j=1,\ldots,p}$ but change the first three eigenvalues of $\boldsymbol{\Lambda}$ to be $15, 10, 5$.

There is leverage effect in Case I. Additionally, γ_t depends on all elements of $\mathbf{W}_t$ and thus Assumption (A.iii) and Assumption (C.iii) are not satisfied. Compared with the first cases, we add spikes for the population eigenvalues in Case II. For more complex situations, we assume that the latent price process are from the factor model for model misspecifications. Assume that

$$d\mathbf{X}_t = \gamma_t^{(1)}\boldsymbol{\Lambda} d\mathbf{W}_t + \gamma_t^{(2)}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{E}^{1/2}d\mathbf{F}_t, \quad t \in [0,1], \tag{4.20}$$

where $\boldsymbol{\beta}$ is a $r \times p$ matrix, $\mathbf{E}$ is a $r \times r$ diagonal covariance matrix, $(\mathbf{F}_t)$ is a r-dimensional standard Brownian motion which is independent of $(\mathbf{W}_t)$ and $\gamma_t^{(1)}, \gamma_t^{(2)}$ are two processes.

- Case III: let the eigenvalues for $\mathbf{E}$ are $(15, 10, 5)$ and the first row of $\boldsymbol{\beta}$ is drown randomly from $U[0.25, 1.75]$ while the remaining two rows are i.i.d $N(0, 0.5^2)$; let $\mathbf{\Lambda} = (0.5^{|i-j|})_{i,j=1,\ldots,p}$, $\gamma_t^{(1)}$ follows (4.19) , $\gamma_t^{(2)}$ follows (4.18) and $(\mathbf{X}_t)$ satisfy (4.20).

Case III follows a factor model and thus Assumption (C.i) is not satisfied. As the following simulation studies show that the proposed estimator even works well under these situations. In all settings, we rescale the spectral of $\mathbf{\Lambda}$ to satisfy $\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}) = p$. We set $n = 23400$ which corresponds one recording every second during 6.5 trading hours and $p = 30$ or 100. $(\boldsymbol{\varepsilon}_t)$ are drawn from independently $N(0, 0.0002\mathbf{I}_p)$. We also compare our method with the data-splitting (DS) in Lam et al. (2017), principal orthogonal complement thresholding (POET) (Aït-Sahalia and Xiu, 2017; Fan et al., 2012). The subsampling interval for Lam et al. (2017) is 15-minute. We use $K = 3$ factors for the principal orthogonal complement thresholding with 0.5 as soft thresholding parameter. To show the advantage of averaging multiple observations, a competitor PATVA-Nonlin is considered based on our proposed estimation procedure without the averaging step described in (2.6). The subsampling interval and pre-averaging window length are both 15-minute. We repeat the simulations 500 times and compare the means of relative spectral loss and relative Frobenius loss for the estimators. For any matrix $\mathbf{A}$, we define the relative spectral and Frobenius loss with respect to ICV matrix as $L_1 = \|\mathbf{A} - \mathbf{\Sigma}\|/\|\mathbf{\Sigma}\|$ and $L_2 = \|\mathbf{A} - \mathbf{\Sigma}\|_F/\|\mathbf{\Sigma}\|_F$, where $\|\mathbf{A}\|$

is the largest singular value of $\mathbf{A}$. Both pre-averaging window length and the subsampling interval are 15-minute. The simulation results are shown in Table 1.

As the results show that the relative losses for the proposed estimator are less than 1 in all cases and the proposed estimator achieves a lot of improvement with respect to the PA-ATVA matrix, which indicate that the nonlinear shrinkage is helpful to the estimation of ICV matrix. From the result for Case II , we can see that the proposed estimator works well even with spikes. More importantly, the proposed estimator is robust with respect to the model misspecification in Case III. Specifically, the proposed estimator endures the leverage effect and factor model. Compared with PATVA-Nonlin, our proposed estimator works better under all circumstances except L_1 loss in Case I. This shows that handling the multiple observations indeed helps account for the market dynamics.

5. Empirical Applications

Now, we study the out-of-sample performance of the proposed method for portfolio allocation. We collected tick-by-tick 30 DJIA stock prices from January 1, 2021 to December 31, 2021, comprising 252 trading days, from the Trade and Quote database. These stocks had excellent liquidity over the

Table 1: Mean estimation-efficiency loss with respect to relative spectral loss and relative Frobenius loss for 6 estimators (RCV, DS, POET, PA-ATVA, PATVA-Nonlin, proposed) for Case I–III. Standard deviation is in bracket. L_1 is the relative spectral loss and L_2 is the relative Frobenius loss. The recording frequency $n = 23400$ and dimension p is set to $p = 30$ or $p = 100$ based on 500 replications.

Methods	Case I		Case II		Case III	
	L_1	L_2	L_1	L_2	L_1	L_2
			$p = 30$			
RCV	5.18(1.5)	4.94(1.35)	1.57(0.43)	2.44(0.63)	2.53(0.44)	3.29(0.53)
DS	1.90(0.63)	2.98(0.85)	0.67(0.15)	1.66(0.37)	0.94(0.2)	1.98(0.33)
POET	5.72(1.66)	4.79(1.33)	1.68(0.48)	2.37(0.63)	2.81(0.46)	3.21(0.53)
PA-ATVA	1.10(0.37)	0.92(0.19)	0.46(0.21)	0.51(0.19)	0.68(0.2)	0.81(0.15)
PATVA-Nonlin	0.66(0.1)	0.74(0.05)	0.54(0.17)	0.61(0.13)	0.66(0.12)	0.72(0.06)
Proposed	0.65(0.15)	0.64(0.08)	0.47(0.17)	0.5(0.14)	0.62(0.13)	0.65(0.07)
			$p = 100$			
RCV	12.20(3.15)	7.83(1.92)	1.55(0.43)	2.79(0.75)	5.81(0.74)	6.52(1.16)
DS	2.93(0.89)	3.35(0.95)	0.66(0.09)	1.48(0.33)	1.38(0.25)	2.77(0.59)
POET	12.88(3.31)	6.21(1.56)	1.57(0.45)	2.28(0.62)	6.14(0.76)	5.19(0.94)
PA-ATVA	3.14(0.55)	1.66(0.16)	0.50(0.19)	0.63(0.16)	1.45(0.31)	1.46(0.16)
PATVA-Nonlin	0.77(0.11)	0.81(0.02)	0.59(0.14)	0.66(0.1)	0.79(0.06)	0.80(0.02)
Proposed	0.88(0.25)	0.76(0.05)	0.52(0.14)	0.57(0.11)	0.74(0.08)	0.75(0.02)

sampling period. We consider the restricted minimum variance portfolio:

$$\min_{\mathbf{w}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}, \quad \text{subject to} \quad \mathbf{w}^{\mathrm{T}}\mathbf{1} = 1 \text{ and } \|\mathbf{w}\|_1 \leq \lambda, \tag{5.21}$$

where $\mathbf{1}$ is a vector of p ones and $\|\mathbf{w}\|_1 = \mathbf{w}^+ + \mathbf{w}^-$ is the gross exposure of portfolio with $\mathbf{w}^+$ being the proportion of long positions and $\mathbf{w}^-$ being the proportion of short positions. When $\lambda = 1$, we consider long only strategies. The global minimum variance portfolio (GMP) achieves when $\lambda = \infty$ and the weight for GMP is $\mathbf{w}_{\mathrm{gmp}} = (\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{1}$ which needs a invertible estimator for ICV matrix. In this section, we plug in different estimators in the construction of minimum variance portfolio problem (5.21). Log prices are averaged within 5 seconds. We consider 15-minute intervals on every trading day from 9:30 to 16:00 for those methods based on "latent" log prices. The pre-averaging window length is also 15-minute. For the PATVA-Nonlin estimator, we apply a 20-minute pre-averaging window length based on every 10-second subsampling observations. Overnight returns and the first 5-minute observations are excluded in the calculation.

We consider 10-day (2-week) training windows and re-evaluate portfolio weights every 5-day (1-week). We compare the out-of-sample performance of the portfolio allocation problem (5.21) with a range of exposure constraints. The global minimum variance portfolio with $\lambda = \infty$ are labeled by GMP. We consider three types of buy and sell strategies.

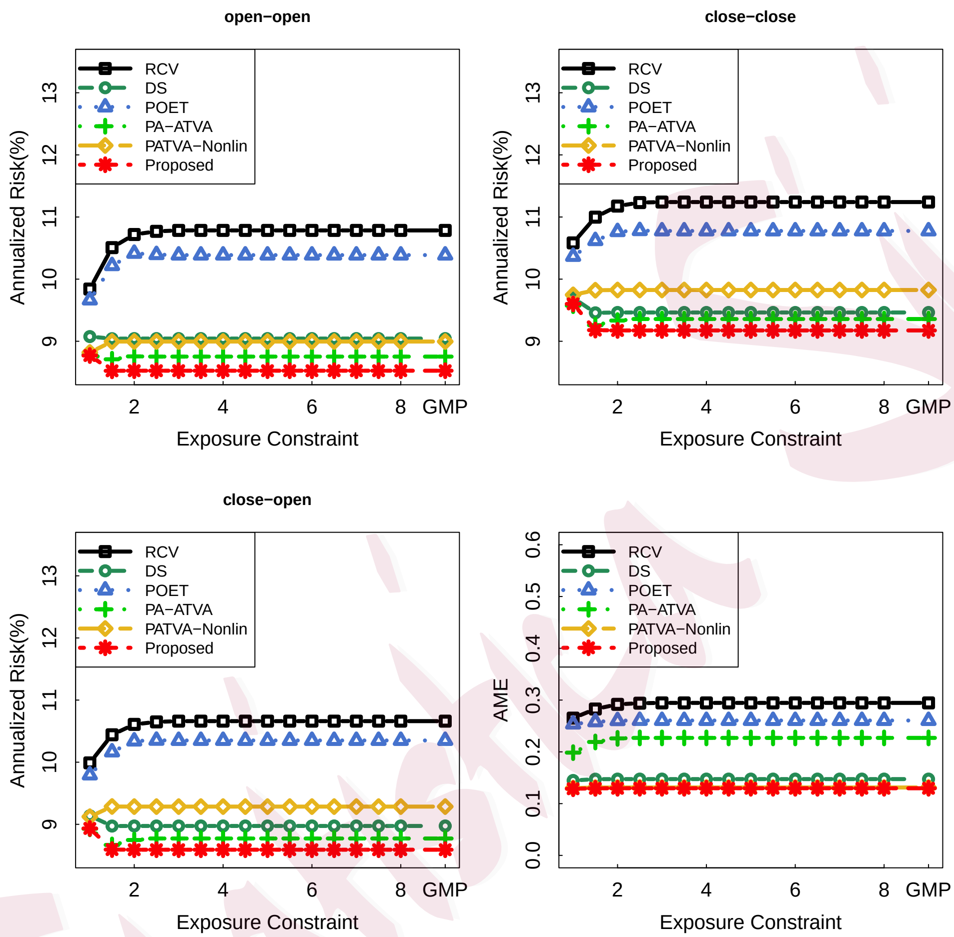

Figure 1: Out-of-Sample risk and average maximum exposures of the portfolio for DJIA index. The x-axis is the exposure constraint λ in the optimization problem (5.21) and the global minimum variance portfolios (GMP) are given at the end of the figures. All abbreviations for different estimators are as in Table 1.

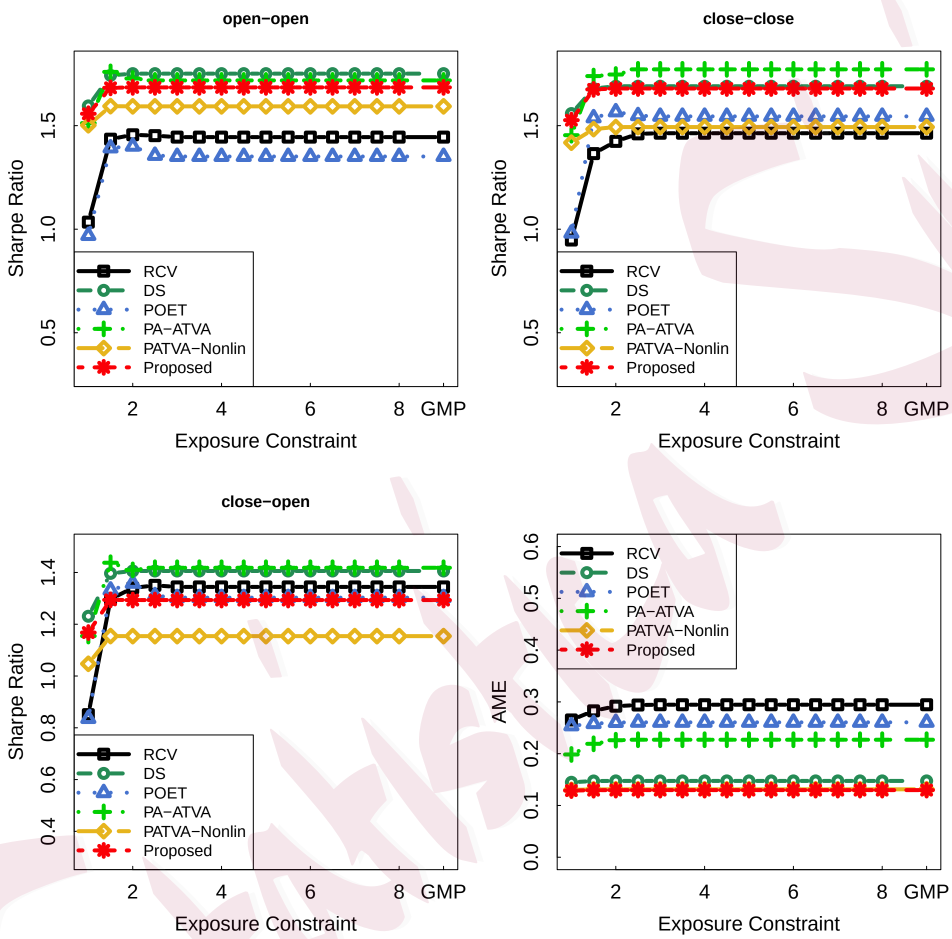

Figure 2: The Sharpe ratio of the portfolio for DJIA index.

- close-open: we assume that investors always buy stocks at the opening of the trading day and sale stocks at the close of the trading day.
- close-close: we assume that investors always buy and sale stocks at the close of the trading day.
- open-open: we assume that investors always buy and sale stocks at the opening of the trading day.

We report the annualized risk and the Sharpe ratio for each strategy and also display the average maximum exposure (AME) which is the the average of the maximum absolute value of $\mathbf{w}$. For risk management, investors may set limitation for the upper bound of the weight for each individual assets. The results are shown in Figures 1 and 2. For comparison, we build up the equal-weight portfolio, which only depends on the size of assets pool and is irrelevant to the exposure constraints. The annualized risks for open-open, close-close, close-open are 10.49 %, 11.26% and 10.54%, respectively.

As the results show that the proposed estimator achieves the smallest GMP and has the smallest maximum exposures. The RCV matrix is mostly sensitive to the tuning parameter λ. The risk for the proposed estimator tends to decrease as λ increasing. It motivates us to use the proposed estimator as an input when predicting the future volatility matrix.

6. Conclusion

This article considers the estimation of ICV matrices based on high-frequency multiple observations. We first look into the limiting spectral property of the averaged version of TVA matrix and find that it works well, i.e., the PA-ATVA matrix effectively eliminates the effects of microstructure noise, multiple transactions, and asynchronous trading. Further, we propose a nonlinear shrinkage estimator and show that it is asymptotically positive-definite and has a good estimation efficiency. As demonstrated by simulated experiments and empirical applications, our proposed estimator has good out-of-sample performance in various contexts. This research is a first step in understanding the impact of multiple transactions in a high-dimensional environment. We leave the study of high-frequency factor models to our future works.

Supplementary Materials

The supplementary material contains some simulation studies and all the proofs of the paper.

Acknowledgements

The authors were supported by NSFC 11871322 and Innovative Research Team of Shanghai University of Finance and Economics 2020110930.

References

Abadir, K. M., W. Distaso, and F. Žikeš (2014). Design-free estimation of variance matrices. *Journal of Econometrics* ***181***(**2**), 165–180.

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

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies* ***18***(**2**), 351–416.

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

Aït-Sahalia, Y. and D. Xiu (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics* ***201***(**1**), 384–399.

Andersen, T. G. and T. Bollerslev (1998). Deutsche mark-dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance* ***53*(1)**, 219–265.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* ***96*(453)**, 42–55.

Bai, Z. D. and Y. Q. Yin (1993). Limit of the smallest eigenvalue of a large-dimensional covariance matrix. *Australian and New Zealand Journal of Statistics* ***21*(3)**, 1275–1294.

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

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics* ***162*(1)**, 149–169.

Barndorff-Nielsen, O. E. and N. Shephard (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *The Annals of Probability* ***64*(2)**, 253–280.

Fan, J., A. Furger, and D. Xiu (2016). Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *Journal of Business & Economic Statistics 34*(4), 489–503.

Fan, J. and D. Kim (2018). Robust high-dimensional volatility matrix estimation for high-frequency factor model. *Journal of the American Statistical Association 113*(523), 1268–

1283. PMID: 30906083.

Fan, J., Y. Li, and K. Yu (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association* ***107***(**497**), 412–428.

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

Jacod, J. and P. Protter (1998). Asymptotic error distributions for the euler method for stochastic differential equations. *Annals of Probability* ***26***(**1**), 267–307.

Karoui, N. E. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics* ***36***(**6**), 2757–2790.

Kim, D., X.-B. Kong, C.-X. Li, and Y. Wang (2018). Adaptive thresholding for large volatility matrix estimation based on high-frequency financial data. *Journal of Econometrics 203*(1), 69–79.

Kim, D., Y. Liu, and Y. Wang (2018). Large volatility matrix estimation with factor-based diffusion model for high-frequency financial data. *Bernoulli 24*(4B), 3657 – 3682.

Kong, X.-B. (2018). On the systematic and idiosyncratic volatility with large panel high-frequency data. *The Annals of Statistics 46*(3), 1077 – 1108.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *The Annals of Statistics* ***44***(**3**), 928–953.

Lam, C., P. Feng, and C. Hu (2017). Nonlinear shrinkage estimation of large integrated covariance matrix. *Biometrika* ***104*(2)**, 481–488.

Ledoit, O. and M. Wolf (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics* ***40*(2)**, 1024–1060.

Li, J. (2013). Robust estimation and inference for jumps in noisy high frequency data: a local-to-continuity theory for the pre-averaging method. *Econometrica* ***81*(4)**, 1673–1693.

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

Tao, M., Y. Wang, Q. Yao, and J. Zou (2011). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *Journal of the American Statistical Association 106*(495), 1025–1040.

Wang, Y. and J. Zou (2010). Vast volatility matrix estimation for high-frequency financial data. *The Annals of Statistics* ***38*(2)**, 943–978.

Xia, N. and X. Zheng (2018). On the inference about the spectra of high-dimensional covariance matrix based on noisy observations-with applications to integrated covolatility matrix inference in the presence of microstructure noise. *The Annals of Statistics* ***46*(2)**, 500–525.

Xiu, D. (2010). Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics* ***159*(1)**, 235–250.

Yin, Y. Q., Z. D. Bai, and P. R. Krishnaiah (1998). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields* ***78*** (**4**), 509–521.

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

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

Zheng, X. and Y. Li (2011). On the estimation of integrated covariance matrices of high dimensional diffusion processes. *Annals of Statistics* ***39*** (**6**), 3121–3151.

Moming Wang

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

wangmoming0902@163.com

Jianhua Hu

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: hu.jianhua@mail.shufe.edu.cn

Ningning Xia

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai

200433, China.

E-mail: xia.ningning@mail.shufe.edu.cn

Yong Zhou

East China Normal University, Shanghai 200062, China.

E-mail: yzhou@amss.ac.cn