ARTICLE IN PRESS

Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Adaptive thresholding for large volatility matrix estimation based on high-frequency financial data

Donggyu Kim [a], Xin-Bing Kong [b], Cui-Xia Li [c], Yazhen Wang [d,*]

[a] *College of Business, Korea Advanced Institute of Science and Technology, Seoul, Republic of Korea*
[b] *School of Science, Nanjing Audit University, Nanjing, Jiangsu, PR China*
[c] *School of Mathematics and Statistics, Lanzhou University, Lanzhou, PR China*
[d] *Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA*

ARTICLE INFO

Article history:
Received 9 January 2017
Received in revised form 26 July 2017
Accepted 27 September 2017
Available online xxxx

JEL classification:
C10
C22
C80

Keywords:
Pre-averaging realized volatility
Regularization
Sparsity
Adaptive thresholding
Diffusion
Integrated volatility

ABSTRACT

Universal thresholding methods have been developed to estimate the large sparse integrated volatility matrix of underlying assets based on high-frequency financial data. Since the integrated volatility matrix often has entries with a wide range of variability, universal thresholding estimators do not take the varying entries into consideration and may have unsatisfactory performances. This paper investigates adaptive thresholding estimation of the large integrated volatility matrix. We first construct an estimator for the asymptotic variance of the pre-averaging realized volatility estimator and then use the two estimators to develop an adaptive thresholding estimator of the large volatility matrix. It is shown that the adaptive thresholding estimator can achieve the optimal rate of convergence over the class of the sparse integrated volatility matrix when both the number of assets and sample size are allowed to go to infinity, while the universal thresholding estimator can achieve only the sub-optimal convergence rate. Also we discuss how to harness the adaptive thresholding scheme in the approximate factor model. The simulation study is conducted to check the finite sample performance of the adaptive thresholding estimators.

1. Introduction

As high-frequency data become available for a wide range of financial assets, researchers have adopted Itô processes to model the log prices of the assets in high-frequency finance, and developed nonparametric methods for estimating their integrated volatility based on high-frequency data contaminated with microstructure noise. For estimating a univariate integrated volatility, popular approaches include realized volatility (Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2002), two-time scale realized volatility (Zhang et al., 2005), multi-scale realized volatility (Zhang, 2006), wavelet realized volatility (Fan and Wang, 2007), pre-averaging realized volatility (Jacod et al., 2009), kernel realized volatility (Barndorff-Nielsen et al., 2008), and a quasi-maximum likelihood estimator (Xiu, 2010). Methods for estimating a multivariate integrated co-volatility consist of multi-scale realized co-volatility based on previous tick data synchronization (Zhang, 2011), a quasi-maximum likelihood estimator based on generalized sampling time (Aït-Sahalia et al., 2010), realized kernel volatility estimator based on refresh time scheme (Barndorff-Nielsen et al., 2011), and pre-averaging realized volatility (Christensen et al., 2010).

Financial research and practices often encounter a large number of assets, and it is well known that the existing multivariate estimation methods do not work well for estimating a large integrated volatility matrix, and in fact they can be inconsistent when both the sample size and the number of assets go to infinity. Regularization approaches such as universal thresholding have been developed for estimating large integrated volatility matrices (see Wang and Zou, 2010; Tao et al., 2013a, b; Kim et al., 2016) under the sparse condition on the integrated volatility matrix. The universal thresholding procedure requires that entries of the integrated volatility matrix are homogeneous. However, volatilities of financial assets usually have entries with a very wide range of variability, which motivates us to develop adaptive thresholding estimator of large volatility matrix with varying entries. We adopt the pre-averaging realized volatility matrix (PRVM) estimator with the generalized sample time and construct estimators for the asymptotic variances of the entries of PRVM estimator. With the asymptotic variance estimator, we select varying thresholds for different entries of the

* Corresponding author.
E-mail addresses: donggyukim@business.kaist.ac.kr (D. Kim), kongxb@fudan.edu.cn (X.-B. Kong), licuixia@lzu.edu.cn (C.-X. Li), yzwang@stat.wisc.edu (Y. Wang).

https://doi.org/10.1016/j.jeconom.2017.09.006
0304-4076/

volatility matrix and call this thresholding estimator the adaptive thresholding estimator. We show that the adaptive thresholding estimator can achieve the optimal convergence rate for the spectral norm under the sparse condition in the asymptotic framework that allows volatility to go to infinity as the sample size and the number of assets go to infinity. Under the same sparse condition, we show that the universal thresholding estimator has the sub-optimal convergence rate. Furthermore, financial markets often exhibit some common market factors such as sector and industry classification, firm size, and price to book ratios. To better accommodate the sparsity for the factor scenario, several estimation methods for factor-based high-dimensional Itô processes have been proposed (Aït-Sahalia and Xiu, 2017; Fan et al., 2016; Kim et al., forthcoming; Kong, 2017a, forthcoming). We also discuss some approach to adopt the adaptive thresholding scheme to the factor model.

The remainder of the paper is organized as follows. Section 2 provides the model and the data structure. Section 3 describes the pre-averaging realized volatility estimator with generalized sampling time and presents the sparse condition and adaptive thresholding estimator. Section 4 establishes their asymptotic behaviors when both the number of variables and the sample size go to infinity. Section 5 features a simulation study to illustrate the finite sample performances of the estimators and applies the adaptive thresholding procedures to a real data set. Section 6 outlines the main steps and key ideas of the proofs, with the supplementary document collecting further detailed technical proofs.

2. The model set-up

Let $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))^T$ be the vector of true log prices of p assets at time t, and assume that $\mathbf{X}(t)$ is an Itô process satisfying,

$$d\mathbf{X}(t) = \boldsymbol{\mu}(t)dt + \boldsymbol{\sigma}(t)^T d\mathbf{B}_t, \quad t \in [0, 1], \tag{1}$$

where $\boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_p(t))^T$ is a drift vector, $\mathbf{B}_t = (B_{1t}, \ldots, B_{pt})^T$ is a standard p-dimensional Brownian motion, and $\boldsymbol{\sigma}(t)$ is a p-by-p matrix. Define the instantaneous volatility of $\mathbf{X}(t)$ as

$$\Sigma(t) = \left(\Sigma_{ij}(t)\right)_{i,j=1,\ldots,p} = \boldsymbol{\sigma}(t)^T \boldsymbol{\sigma}(t),$$

and its quadratic variation,

$$[\mathbf{X}, \mathbf{X}]_t = \int_0^t \Sigma(s)ds = \left(\int_0^t \Sigma_{ij}(s)ds\right)_{i,j=1,\ldots,p}, \quad t \in [0, 1].$$

In high-frequency finance, the true log prices $\mathbf{X}(t)$ are observed with micro-structure noises, and the high-frequency prices of different assets are recorded at mismatched time points, which is called a non-synchronization problem. In light of these, we assume that observed high-frequency financial data $Y_i(t_{i,l})$ obey the model,

$$Y_i(t_{i,l}) = X_i(t_{i,l}) + \epsilon_i(t_{i,l}), \quad i = 1, \ldots, p, l = 0, \ldots, n_i, \tag{2}$$

where $t_{i,l}$ denotes the lth observation time point for the ith asset, $\epsilon_i(t_{i,l})$, $i = 1, \ldots, p$, $l = 0, \ldots, n_i$, are independent noises with mean zero, for each fixed i, $\epsilon_i(t_{i,l})$, $l = 0, \ldots, n_i$, are i.i.d. random variables with variance η_{ii}, and $\epsilon_i(\cdot)$ and $X_i(\cdot)$ are independent.

The goal of this paper is to adaptively estimate the large volatility matrix

$$\Gamma = \left(\Gamma_{ij}\right)_{i,j=1,\ldots,p} = [\mathbf{X}, \mathbf{X}]_1,$$

and investigate the asymptotic properties of the proposed volatility matrix estimators in the framework that allows both the number of assets and the sample size to go to infinity.

3. Large volatility matrix estimation

3.1. Pre-averaging realized volatility matrix

To construct a realized co-volatility matrix for multiple assets based on non-synchronized and noisy high-frequency financial data, we first need some scheme to synchronize observation time points. In this paper, we adopt the generalized sampling time scheme (Aït-Sahalia et al., 2010) which may include other available time synchronization schemes such as refresh time (Barndorff-Nielsen et al., 2011) and previous tick (Wang and Zou, 2010; Zhang, 2011).

Definition 1 (*Aït-Sahalia et al., 2010*)**.** A sequence of time points $\boldsymbol{\tau} = \{\tau_0, \ldots, \tau_n\}$ is said to be the generalized sampling time for a collection of p assets, if

(1) $0 = \tau_0 < \tau_1 < \tau_2 \cdots < \tau_n = 1$;
(2) the time intervals, $\{\Delta_j^\tau = \tau_j - \tau_{j-1}, 1 \le j \le n\}$, satisfy $\sup_j \Delta_j^\tau \xrightarrow{p} 0$;
(3) there exists at least one observation for each asset between consecutive τ_i's.

For asset i, with the generalized sampling time scheme, we select an arbitrary observation, $Y_i(\tau_{i,l})$, between τ_{l-1} and τ_l, that is, $\tau_{i,l} \in (\tau_{l-1}, \tau_l] \cap \{t_{i,k}, k = 0, 1, \ldots, n_i\}$, $i = 1, \ldots, p$.

To deal with the micro-structure noise, we use the pre-averaging realized volatility estimators (Jacod et al., 2009; Christensen et al., 2010).

Definition 2 (*Jacod et al., 2009; Christensen et al., 2010*)**.** For the generalized sampling time $\boldsymbol{\tau}$ in Definition 1, the pre-averaging realized volatility matrix (PRVM) estimator is given by

$$\widehat{\Gamma} = \frac{1}{\psi K} \sum_{l=1}^{n-K+1} \left[\bar{\mathbf{Y}}(\tau_l)\bar{\mathbf{Y}}(\tau_l)^T - \varsigma \widehat{\boldsymbol{\eta}}\right],$$

where

$$\bar{\mathbf{Y}}(\tau_i) = \sum_{l=1}^{K-1} g\left(\frac{l}{K}\right) \left[\mathbf{Y}(\tau_{i+l}) - \mathbf{Y}(\tau_{i+l-1})\right],$$

$$\varsigma = \sum_{l=0}^{K-1} \left[g\left(\frac{l}{K}\right) - g\left(\frac{l+1}{K}\right)\right]^2, \quad \psi = \int_0^1 g(t)^2 dt,$$

$$\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_{ij})_{i,j=1,\ldots,p} = \mathrm{diag}(\widehat{\eta}_1, \ldots, \widehat{\eta}_p),$$

$$\widehat{\eta}_i = \frac{1}{2n_i} \sum_{l=1}^{n_i} \left[Y_i(t_{i,l}) - Y_i(t_{i,l-1})\right]^2,$$

with $K = \varphi_k n^{1/2}$ for some constants φ_k, and the weight function $g(\cdot)$ is continuous and piecewise continuously differentiable with a piecewise Lipschitz derivative g' and $g(0) = g(1) = 0$, and satisfies $\int_0^1 g^2(t)dt > 0$.

3.2. Adaptive thresholding estimators

Assume that the integrated volatility matrix Γ belongs to a sparse class

$$\mathcal{F}_\delta(\pi(p)) = \left\{\Gamma : \Gamma \succ 0, \max_{1 \le i \le p} \sum_{j=1}^p (\Gamma_{ii}\Gamma_{jj})^{(1-\delta)/2} \left|\Gamma_{ij}\right|^\delta \right. \\ \left. \le \Theta\pi(p) \text{ and } E\left[\Theta^4\right] \le C_\Theta \right\}, \tag{3}$$