Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications 126 (2016) 3527–3577

stochastic processes and their applications

www.elsevier.com/locate/spa

Asymptotic theory for large volatility matrix estimation based on high-frequency financial data

Donggyu Kim[a,*], Yazhen Wang[a], Jian Zou[b]

[a] *Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA*
[b] *Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280, USA*

Received 19 September 2015; received in revised form 9 April 2016; accepted 9 May 2016
Available online 14 May 2016

Abstract

In financial practices and research studies, we often encounter a large number of assets. The availability of high-frequency financial data makes it possible to estimate the large volatility matrix of these assets. Existing volatility matrix estimators such as kernel realized volatility and pre-averaging realized volatility perform poorly when the number of assets is very large, and in fact they are inconsistent when the number of assets and sample size go to infinity. In this paper, we introduce threshold rules to regularize kernel realized volatility, pre-averaging realized volatility, and multi-scale realized volatility. We establish asymptotic theory for these threshold estimators in the framework that allows the number of assets and sample size to go to infinity. Their convergence rates are derived under sparsity on the large integrated volatility matrix. In particular we have shown that the threshold kernel realized volatility and threshold pre-averaging realized volatility can achieve the optimal rate with respect to the sample size through proper bias corrections, but the bias adjustments cause the estimators to lose positive semi-definiteness; on the other hand, in order to be positive semi-definite, the threshold kernel realized volatility and threshold pre-averaging realized volatility have slower convergence rates with respect to the sample size. A simulation study is conducted to check the finite sample performances of the proposed threshold estimators with over hundred assets.

Keywords: Multi-scale realized volatility; Kernel realized volatility; Pre-averaging realized volatility; Regularization; Sparsity; Threshold; Diffusion; Integrated volatility

* Corresponding author.
E-mail addresses: kimd@stat.wisc.edu (D. Kim), yzwang@stat.wisc.edu (Y. Wang), jzou@wpi.edu (J. Zou).

http://dx.doi.org/10.1016/j.spa.2016.05.004
0304-4149/

1. Introduction

Volatility analysis for high-frequency financial data is a vibrant research area in financial econometrics and statistics. With high-frequency data, we can better study market microstructure and directly estimate market volatility. Several volatility estimation methods have been developed. Estimators of a univariate integrated volatility include realized volatility (RV) [3,6], two-time scale realized volatility (TSRV) [30], multi-scale realized volatility (MSRV) [28], wavelet realized volatility (WRV) [15], kernel realized volatility (KRV) [4], pre-averaging realized volatility (PRV) [17], and a quasi-maximum likelihood estimator (QMLE) [27]. For multiple assets, we encounter a non-synchronization problem caused by the fact that transactions for different assets occur at distinct time, and the high-frequency data are observed at mismatched time points. Methods for estimating an integrated co-volatility consist of multi-scale realized co-volatility based on previous tick data synchronization [29], realized kernel volatility estimator based on refresh time scheme [5], a quasi-maximum likelihood estimator based on generalized sampling time [1], and pre-averaging realized volatility [12] (see also [7,13,19–21,25]). These estimators have good performances for a relatively small number of assets. When there are a large number of assets in financial practices such as asset pricing, portfolio allocation, and risk management, the volatility estimators designed for estimating a small integrated volatility matrix perform very poorly, and in fact, they are inconsistent when both the number of assets and sample size go to infinity [26]. For the case of a large number of assets, we need to impose some sparse structure on the integrated volatility matrix and employ regularization such as thresholding to obtain consistent estimators of the large volatility matrix [22,24,23]. In particular, Tao et al. [22,24] investigated convergence rates of multi-scale realized volatility matrix estimator in the asymptotic framework that allows both the number of assets and sample size to go to infinity, and showed that the estimator achieves optimal convergence rate with respect to the sample size.

This paper considers the kernel realized volatility (KRV) [5,4], the pre-averaging realized volatility (PRV) [12,17], and the multi-scale realized volatility (MSRV) [22,29] based on generalized sampling time scheme. We investigate their convergence rates in the asymptotic framework that both the number of assets and sample size go to infinity. Our asymptotic analyses are under dependent micro-structure noises, finite moment conditions on asset prices and sparsity on the integrated volatility matrix. We have shown that the estimators based on MSRV, KRV, and PRV with proper bias corrections can achieve the optimal convergence rate with respect to the sample size, but they may not be positive semi-definite. In order to be positive semi-definite, KRV and PRV estimators can only achieve a slower convergence with respect to the sample size. There is a trade-off between positive semi-definiteness and fast convergence rate.

The rest of the paper is organized as follows. Section 2 provides the price model and the data structure. Section 3 describes five realized volatility estimators based on the kernel realized volatility, pre-averaging realized volatility, and multi-scale realized volatility with generalized sampling time. Section 4 presents a sparse condition and regularized estimators and establishes their asymptotic behaviors when both the number of assets and the sample size go to infinity. Section 5 features a simulation study to illustrate the finite sample performances of the estimators. Section 6 outlines the key ideas and main steps of the proofs, with Appendix collected further detailed technical proofs.

2. The model set-up

Denote by $\mathbf{X}(t) = (X_1(t), \dots, X_p(t))^T$ the vector of true log prices of p assets at time t. Modern finance theory usually assumes that $\mathbf{X}(t)$ follows a continuous-time diffusion