Journal of Econometrics 180 (2014) 217–232

Contents lists available at ScienceDirect

Journal of Econometrics

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

A quasi-maximum likelihood approach for integrated covariance matrix estimation with high frequency data

Cheng Liu [a,b], Cheng Yong Tang [c,*]

[a] Economics and Management School of Wuhan University, Hubei, China

[b] Sim Kee Boon Institute for Financial Economics, Singapore Management University, Singapore

[c] Business School, University of Colorado Denver, United States

ARTICLE INFO

Article history:
Received 5 October 2012
Received in revised form
31 January 2014
Accepted 31 January 2014
Available online 27 February 2014

JEL classification:
C13
C22
C51

Keywords:
High frequency data
Integrated covariance matrix
Microstructure noises
Quasi-maximum likelihood

ABSTRACT

Estimating the integrated covariance matrix (ICM) from high frequency financial trading data is crucial to reflect the volatilities and covariations of the underlying trading instruments. Such an objective is difficult due to contaminated data with microstructure noises, asynchronous trading records, and increasing data dimensionality. In this paper, we study a quasi-maximum likelihood (QML) approach for estimating an ICM from high frequency financial data. We explore a novel multivariate moving average time series device that is convenient for evaluating the estimator both theoretically for its asymptotic properties and numerically for its practical implementations. We demonstrate that the QML estimator is consistent to the ICM, and is asymptotically normally distributed. Efficiency gain of the QML approach is theoretically quantified, and numerically demonstrated via extensive simulation studies. An application of the QML approach is illustrated through analyzing a high frequency financial trading data set.

1. Introduction

The integrated covariance matrix (ICM) of multiple financial returns over a period of time plays a crucial role in risk management and many financial applications including constructing hedging and investing strategies, pricing stock options and other derivatives etc. Contemporary high frequency trading data become increasingly available and have attracted substantial recent interests for estimating volatilities and covariations between financial returns; see, for example, Zhang et al. (2005), Aït-Sahalia et al. (2005, 2010) and references therein.

Features of high frequency trading data such as the contamination by microstructure noises and asynchronous trading times make it challenging to estimate volatilities and covariations (Aït-Sahalia et al., 2005, 2010). Methods have been extensively developed to deal with contaminated data when estimating the univariate integrated volatility of an individual asset—see, for example, the realized kernel approach (Barndorff-Nielsen et al., 2008), the two and multiple time scale methods (Zhang et al., 2005; Zhang, 2006), the power of variation approaches (Barndorff-Nielsen and Shephard, 2004), the wavelet methods dealing with jumps (Fan and Wang, 2007), the pre-averaging approach (Jacod et al., 2009), and the quasi-maximum likelihood (QML) approach (Aït-Sahalia et al., 2005; Xiu, 2010).

Clearly, extending the univariate methods for volatilities to multivariate cases for the integrated covariance matrix is substantially more challenging because jointly dealing with the microstructure noises and asynchronous data among multivariate return series is more difficult. Methods have recently been developed for estimating the ICM—see, for example, the multivariate realized kernels approach (Barndorff-Nielsen et al., 2011), the pre-averaging approach (Christensen et al., 2010), the two time scales method (Zhang, 2011), the threshold average realized volatility matrix approach (Wang and Zou, 2010), and the frequency domain approaches based on the Fourier transformations (Malliavin and Mancino, 2002, 2009). Recently, Aït-Sahalia et al. (2010) propose to estimate the ICM via pairwise covariances estimations. Their approach initially estimate univariate integrated variances of constructed new series using the univariate quasi-maximum likelihood approach, and then estimate the corresponding pairwise covariances using a polarization identity. The quasi-maximum

* Correspondence to: Campus box 165, PO Box 173364, Denver, CO 80217-3364, United States.

E-mail address: chengyong.tang@ucdenver.edu (C.Y. Tang).

http://dx.doi.org/10.1016/j.jeconom.2014.01.008
0304-4076/

likelihood approach for a univariate volatility is optimal if the data model is correctly specified. However, a subsequent covariance estimation by using the polarization identity may not adequately incorporate all data information. In addition, pairwise estimations of covariances cannot guarantee the estimated ICM to be non-negative definite.

In this paper, we study a quasi-maximum likelihood (QML) approach for estimating the ICM with high frequency financial data. Univariate QML approach is originated in Aït-Sahalia et al. (2005) where constant volatility is assumed, and Xiu (2010) studies the properties the QML approaches for integrated variance estimation with stochastic volatility. However, extending the QML approach to multivariate ICM estimation is difficult in both practical implementations and theoretical analysis because a daunting huge covariance matrix is encountered. Our study makes the following contributions. First, to overcome the difficulties of the QML approach in practical implementation and theoretical analysis, we discover a novel and convenient device that constructs a huge covariance matrix from a much lower dimensional multivariate moving average time series model. Such a device itself is interesting and can be further explored for extensively studying the ICM estimation. Second, we explore the use of the QML approach for more adequately incorporate data information, and demonstrate the gain in efficiency. We show that the QML approach can handle microstructure noises, which successfully extends the QML approach in univariate cases to multivariate ones. Consistency, efficiency and robustness are also achieved by the QML approach. We note that the QML approach estimates all covariations jointly over its parameter space so that the estimated ICM is consistent to a positive definite limit. On the other hand, positive definiteness of the estimator can also be conveniently ensured in practical implementations by using an appropriate parameterization of the covariance matrix in the quasi-likelihood function via using the Cholesky type decompositions.

We begin our study with synchronous data for illustrating the QML approach for simplicity and clarity in demonstrating its theoretical analysis and practical implementation. Then we show that the QML approach remains consistent for asynchronous data scenario with appropriate synchronization. For asynchronous data, we note that existing synchronization strategies such as the previous tick approach (Zhang, 2011), the refresh time scheme (Barndorff-Nielsen et al., 2011), MINSPAN (Harris et al., 1995), and the generalized synchronization method (Aït-Sahalia et al., 2010) can be applied for pre-processing the data.

We would like to note two concurrent and independent works of our paper—Shephard and Xiu (2012) and Corsi et al. (2012) for ICM estimation with high frequency financial data. Both Shephard and Xiu (2012) and Corsi et al. (2012) apply the expectation–maximization (EM) algorithm for dealing with asynchronous and irregularly spaced data, and Shephard and Xiu (2012) comprehensively investigate the properties of the QML approach under various scenarios for bivariate observations. We would like to single out the difference and connection of our study with Shephard and Xiu (2012) as follows. First, our investigation is based on the novel investigation of the multivariate moving average time series model approach that naturally extends the univariate case as in Aït-Sahalia et al. (2005) and Xiu (2010). While in Shephard and Xiu (2012), derivations based on the original data model are developed. As shown in our technical development (Proposition 1 and Lemma 2), a very convenient eigendecomposition with well structured matrices can be established for analyzing the QML approach. Therefore, results for general multivariate returns with the QML approach can be developed with general assumptions on the correlations among the noises. With more details in a later section, the finding of our study with bivariate data is consistent with those in Shephard and Xiu (2012) that are independently developed using different techniques. In addition, our study shows that the connection to the multivariate time series model also facilitates the construction of an efficient algorithm for evaluating and optimization the quasi-likelihood function.

The rest of this paper is structured as follows. We describe the proposed approach and the novel device for theoretical analysis and practical implementation in Section 2. Main results are given in Section 3, and followed by simulations and an example of high frequency financial data analysis in Section 4. We conclude the paper with some discussions in Section 5.

2. Methodology

We denote by $\tilde{\mathbf{Y}}_t = (\tilde{Y}_{1t}, \tilde{Y}_{2t}, \ldots, \tilde{Y}_{dt})'$ the observed log-prices of d assets at time $t \in [0, T]$ for a fixed T. Without loss of generality, we take $T = 1$ for simplicity hereinafter. Suppose that each $\tilde{Y}_{it}$ $(i = 1, \ldots, d)$ contains the true log-price X_{it} and microstructure noise U_{it}—i.e., $\tilde{Y}_{it} = X_{it} + U_{it}$. We impose the following assumption in our study.

Assumption 1. The true log-price process $\mathbf{X}_t = (X_{1t}, \ldots, X_{dt})'$ satisfies:

$$dX_{it} = \mu_{it} dt + \sigma_{it} dW_{it} \quad \text{and}$$
$$E(dW_{it} dW_{kt}) = \rho_{ikt} dt \quad (i, k = 1, \ldots, d),$$

where each drift process μ_{it} is locally bounded and the spot volatility process σ_{it} is positive and locally bounded Itô semimartingale, $W_{1t}, \ldots, W_{dt}$ are univariate Brownian motions, and $\rho_{ikt} = \frac{E(W_{it} W_{kt})}{t}$ $(i, k = 1, \ldots, d)$.

Here following Aït-Sahalia et al. (2010), correlations among the latent log-prices are introduced by correlated Brownian motions. We note that the impact on estimating integrated volatilities due to μ_{it} is asymptotically negligible when sampling interval lengths shrink to zero in high frequency financial data analysis (Mykland and Zhang, 2012). Thus for simplicity and without compromising the general validity of our method, we assume $\mu_{it} = 0$ hereinafter.

Our goal is to estimate the ICM of the log-price $\mathbf{X}_t$:

$$\mathbf{\Sigma_I} = \int_0^1 \mathbf{\Sigma}_t dt$$

where $\mathbf{\Sigma}_t = (\sigma_{it}\sigma_{kt}\rho_{ikt})_{i,k=1}^d$ is the instantaneous covariance matrix of $\mathbf{X}_t$. Hence, the diagonal elements in $\mathbf{\Sigma_I}$ reflect the integrated volatilities and the other elements reflect the integrated covariations between assets. For our theoretical analysis, we first consider synchronous data observed at equally spaced time points on [0, 1]. We denote by $\Delta = n^{-1}$ the sampling interval where n is the sample size of the high frequency data.

The QML approach for the ICM estimation with no microstructure noise – i.e., $\mathbf{X}_t$ is directly observed – is straightforward to carry out by examining the log-return $\mathbf{Y}_{t_j} = \mathbf{X}_{t_j} - \mathbf{X}_{t_{j-1}}$. For simplicity in notations, we suppress the time t in the index and treat $\mathbf{Y}_{t_j}$ as $\mathbf{Y}_j$, $\mathbf{X}_{t_j}$ as $\mathbf{X}_j$ when no confusion arises. One can always impose a not necessarily correct model by assuming that $\mathbf{Y}_j$ $(j = 1, \ldots, n)$ independently follow a multivariate normal distribution $N(0, \mathbf{\Sigma}\Delta)$ where $\mathbf{\Sigma}$ is a time invariant covariance matrix. Then the quasi-log-likelihood function is

$$l(\mathbf{\Sigma}) = -\frac{n}{2}\log\det(\mathbf{\Sigma}\Delta) - \frac{nd}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{n}\mathbf{Y}_j'(\mathbf{\Sigma}\Delta)^{-1}\mathbf{Y}_j.$$

Subsequently, the QML estimator of $\mathbf{\Sigma}$ is given by

$$\hat{\mathbf{\Sigma}} = \sum_{j=1}^{n}\mathbf{Y}_j\mathbf{Y}_j',$$