Statistics and Its Interface Volume 6 (2013) 463–475

A state space model approach to integrated covariance matrix estimation with high frequency data*

Cheng Liu and Cheng Yong Tang†

We consider a state space model approach for high frequency financial data analysis. An expectation-maximization (EM) algorithm is developed for estimating the integrated covariance matrix of the assets. The state space model with the EM algorithm can handle noisy financial data with correlated microstructure noises. Difficulty due to asynchronous and irregularly spaced trading data of multiple assets can be naturally overcome by considering the problem in a scenario with missing data. Since the state space model approach requires no data synchronization, no record in the financial data is deleted so that it efficiently incorporates information from all observations. Empirical data analysis supports the general specification of the state space model, and simulations confirm the efficiency gain and the benefit of the state space model approach.

Keywords and phrases: EM algorithm, High frequency data, Integrated covariance matrix, Kalman Filter, Microstructure noise, Missing data, Quasi-maximum likelihood, State Space Model.

1. INTRODUCTION

Since the seminal work of Engle (1982), investigating variations and covariations among multivariate time series of assets prices has been a central focus of both quantitative and empirical financial studies; see Andersen, Bollerslev, Diebold, and Lays (2003), Andersen, Bollerslev, and Diebold (2008), Barndorff-Nielsen and Shephard (2007) and reference therein for comprehensive overviews. In the past two decades, high frequency financial trading data have become increasingly available, and there are surging research interests covering both volatility estimation for univariate return series, and covariations estimation among assets; see, for example, Barndorff-Nielsen and Shephard (2004), Aït-Sahalia, Mykland, and Zhang (2005), Zhang, Mykland, and Aït-Sahalia (2005), Hansen and Lunde (2006), Fan and Wang (2007), Barndorff-Nielsen, Hansen, Lunde, and Shepard (2008,2011), Aït-Sahalia, Fan, and Xiu (2010), and Christensen, Kinnebrock, and Podolskij (2010).

The abundance of high frequency financial data has blessed the investigations of volatilities and covariations for the underlying price processes over a short period of time to more precisely reflect the current market dynamics. Thus it can be viewed as more advantageous compared with traditional volatility and covariations modeling and forecasting approaches that require observations over a longer period of time. On the other hand, however, features of high frequency financial data also pose new challenges and difficulties. First of all, it is common that observed financial trading data are contaminated by the so-called market microstructure noises. The impact of data contamination on the volatilities and covariations estimations for the unobservable underlying price processes is very substantial, especially in studies using high frequency data for summarizing market dynamics over a short period of time; see, for example, Hansen and Lunde (2006) for an overview. Additional difficulty arises from practical features of the financial trading data including the so-called asynchronous and irregularly time spaced observations; see, for example, discussions in Barndorff-Nielsen, Hansen, Lunde, and Shepard (2011), and Aït-Sahalia, Fan, and Xiu (2010). The last but not the least challenge is due to the large number of assets of interest so that the problem of volatility and covariations estimation belongs to the well known family of difficult problems of estimating a huge covariance matrix; see Wang and Zhou (2011), Tao, Wang, Yao and Zou (2011) among others.

Estimating univariate realized volatility from high frequency trading data is influenced by the main difficulty due to contaminated data; see Hansen and Lunde (2006) and Aït-Sahalia, Mykland, and Zhang (2005) among others. Various methods have been developed to deal with contaminated data and are demonstrated effective for estimating the integrated volatility. These include the realized kernel approach (Barndorff-Nielsen, Hansen, Lunde, and Shepard, 2008), the two and multiple time scale approach (Zhang, Mykland, and Aït-Sahalia, 2005; Zhang, 2006; Aït-Sahalia and Mykland, 2009), pre-averaging approach (Jacod, Li, Mykland, Podolskij, and Vetter, 2009),

* We thank Professor Yazhen Wang for insightful suggestions that have improved the presentation of the paper. The main work of this paper was done when Liu was a graduate student at the Department of Statistics and Applied Probability, National University of Singapore. Research support from the National University of Singapore is gratefully acknowledged. Tang acknowledges research support from the Business School, University of Colorado Denver.

†Corresponding author.

and the quasi-maximum likelihood approach (Aït-Sahalia, Mykland, and Zhang, 2005; Xiu, 2010).

Dealing with multivariate assets requires extra effort when estimating covariations because the observations of trading prices are generally asynchronous among the assets. A class of methods in this scenario is pre-processing the data set by applying a variety of synchronizing schemes such as the previous tick approach (Zhang, 2011), the fresh time scheme (Barndorff-Nielsen, Hansen, Lunde, and Shepard, 2011) and the MINSPAN (Harris, McInish, Shoesmith, and Wood, 1995), and the Generalized Synchronization method (Aït-Sahalia, Fan, and Xiu, 2010). Subsequently, methods developed for synchronized high frequency data can be applied; see, for example, the realized kernel approach (Barndorff-Nielsen, Hansen, Lunde, and Shepard, 2011), the pre-averaging approach (Christensen, Kinnebrock, and Podolskij, 2010), the two time scale method (Zhang, 2011), the threshold average realized volatility matrix method (Wang and Zou, 2010), and the quasi-maximum likelihood approach (Aït-Sahalia, Fan, and Xiu, 2010; Liu and Tang, 2012). It is clear that those data synchronizing methods inevitably delete a portion of the observations, and thus efficiency loss may incur in the estimators. Another class of methods dealing with asynchronous data is by inserting pseudo-data into the original data set by some interpolations before applying the aforementioned methods; see, for example, Hoshikawa, Kanatani, Nagai and Nishiyama (2008), Peluso, Corsi, and Mira, 2012, and Malliavin and Mancino (2002, 2009). Inserting data may induce bias in the estimators because there is no guarantee that a data interpolation method can accurately reflect the properties of the unknown data model.

In this paper, we consider a state space model approach for studying multivariate contaminated high frequency financial data that can be observed asynchronously over irregularly spaced times. By considering asynchronous trading data in the scenario of missing data with incomplete observations, an expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) is developed to jointly estimate the volatilities and covariations of the underlying asset prices in a covariance matrix. Our study of high frequency financial data analysis using a state space model adapts and generalizes the multivariate quasi-maximum likelihood approach of Liu and Tang (2012) for more conveniently dealing with general asynchronous and irregularly spaced data. Instead of using data synchronization methods for pre-processing asynchronous data, the state space model approach is able to handle the original data directly. The state space model approach shares the same data model with the one in the quasi-maximum likelihood approach, and they are equivalent if all data are observed synchronously. We show the state space model approach is convenient for practical implementations, and is capable of efficiently incorporating data information without manipulating the original data by deleting or inserting records. Our simulation studies demonstrate the efficiency gain by using the proposed approach.

We note two independent studies, Shepard and Xiu (2012) and Crosi, Peluso, and Audrino (2012), on high frequency financial data analysis using the EM algorithm. In both Shepard and Xiu (2012) and Crosi Peluso, and Audrino (2012), the microstructure noises are considered as uncorrelated between different assets. As detailed in Section 2, we develop a more general EM algorithm for the state space model approach that allows the correlations among the microstructure noises to take a general form. As shown in our simulation studies, there is a substantial impact on the estimation of the integrated covariance matrix if the structure covariance matrix of the microstructure noise is misspecified. Our empirical financial data analysis also reveals that it is more reasonable to consider the covariance matrix of the microstructure noise to be a general positive definite matrix.

The rest of this paper is structured as follows. We describe the proposed state space model approach in Section 2. Simulations and an example of high frequency financial data analysis are presented in Section 3, and Section 4 concludes the paper.

2. THE STATE SPACE MODEL APPROACH

Let us introduce some notations first. We denote by $\mathbf{Y}_t = (Y_{1t}, Y_{2t}, \ldots, Y_{dt})'$ the observed log-prices of d assets at time t over a fixed interval $[0, T]$. Without loss of generality, we take $T = 1$ for simplicity hereinafter. Suppose that each Y_{it} $(i = 1, \ldots, d)$ contains the true log-price X_{it} and microstructure noise U_{it} with the additive form $Y_{it} = X_{it} + U_{it}$.

The true log-price process $\mathbf{X}_t = (X_{1t}, \ldots, X_{dt})'$ are assumed to satisfy:

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t d\mathbf{W}_t, \tag{1}$$

where the drift process $\boldsymbol{\mu}_t$ is assumed to be locally bounded and spot volatility process $\boldsymbol{\sigma}_t$ is positive and locally bounded Itô semimartingale matrix, $\mathbf{W}_t = (W_{1t}, \ldots, W_{dt})'$ is an independent d dimensional Brownian motion, $(\rho_{klt})_{kl} := (\frac{E(W_{kt}W_{lt})}{t})_{1\le k,l\le d}$ is a positive definite correlation matrix.

For high frequency financial data analysis, the impact due to the mean $\boldsymbol{\mu}_t$ is asymptotically negligible when sampling interval lengths shrink to zero if $\boldsymbol{\mu}_t$ is locally bounded (Mykland and Zhang, 2010). Thus we consider for simplicity that $\boldsymbol{\mu}_t = \mathbf{0}$. Our target is to estimate the integrated covariance matrix (ICM) of the log-price $\mathbf{X}_t$:

$$\boldsymbol{\Sigma}_{\mathbf{I}} = \int_0^1 \boldsymbol{\sigma}_t \boldsymbol{\sigma}_t' dt = \int_0^1 \boldsymbol{\Sigma}_t dt.$$

The d-dimensional noises $\mathbf{U}_t = (U_{1t}, \ldots, U_{dt})'$ contaminated in different observations of $\mathbf{Y}_t$ are typically assumed to be independent and identically distributed with mean $\mathbf{0}$, positive definite covariance matrix $\mathbf{A}_0$ and finite fourth moment; see, for example, Aït-Sahalia, Fan, and Xiu (2010) and

Liu and Tang (2012). In addition, $\mathbf{U}_t$ and $\mathbf{X}_t$ are assumed to be mutually independent to ensure the identifiability of the ICM $\mathbf{\Sigma_I}$ and $\mathbf{A}_0$. For discussions about the impact of serially-correlated noises, we refer to Aït-Sahalia and Mykland (2009) and Aït-Sahalia, Mykland, and Zhang (2011).

In our study, the observations of the assets price processes are allowed to be asynchronous and irregularly spaced over $[0,1]$. Therefore, we denote the collection of data by $\{Y_{it_{ij}}, i = 1,\ldots,d; j = 1,\ldots,n_i\}$ where $Y_{it_{ij}}$ denotes the observation for the ith asset at time t_{ij} for $j = 1,\ldots,n_i$ with t_{ij} and n_i being asset specific.

When t_{ij} are synchronous for all assets and equally spaced with interval Δ over $[0,1]$ for all assets, Liu and Tang (2012) analyze the properties of a quasi maximum likelihood approach that imposes two not necessarily correct assumptions that a) $\boldsymbol{\sigma}_t = \boldsymbol{\sigma}$ in (1) so that $\mathbf{\Sigma}_t = \mathbf{\Sigma}$ is time invariate, and b) $\mathbf{U}_t$ is a normally distributed random vector and independent of $\mathbf{X}_t$. They show that the estimators of $\mathbf{\Sigma}$ and $\mathbf{A}$ of the quasi-maximum likelihood approach are consistent to the ICM $\mathbf{\Sigma_I}$ and $\mathbf{A}_0$ as $\Delta \to 0$. Moreover, the estimator of the quasi-maximum likelihood (QML) approach achieves the optimal rate of convergence in the sense of Gloter and Jacod (2001).

To extend the quasi-maximum likelihood approach by relaxing the requirement on synchronous data, we consider the following state space model. We firstly write the union of all observation time points t_{ij}s $(i = 1,\ldots,d; j = 1,\ldots,n_i)$ of d assets as

$$\tau_j, j = 1,\ldots,n,$$

where n is the total number of distinct time points that each one has observations of at least one asset price, and τ_js are those observations times such that $0 \le \tau_1 < \tau_2 < \cdots < \tau_j < \cdots < \tau_n \le 1$. We follow the settings in Liu and Tang (2012) and impose the aforementioned two not necessarily correct assumptions a) and b). By ignoring the impact of $\boldsymbol{\mu}_t$ in (1) and since hypothetically the underlying process can be observed at any time, we have

$$\begin{aligned}\mathbf{X}_{\tau_j} - \mathbf{X}_{\tau_{j-1}} &= \boldsymbol{\sigma}(\mathbf{W}_{\tau_j} - \mathbf{W}_{\tau_{j-1}}) \sim N(0, \mathbf{\Sigma}\Delta_j),\\ \mathbf{Y}_{\tau_j} - \mathbf{Y}_{\tau_{j-1}} &= \mathbf{X}_{\tau_j} - \mathbf{X}_{\tau_{j-1}} + \mathbf{U}_{\tau_j} - \mathbf{U}_{\tau_{j-1}}\\ &\sim N(0, \mathbf{\Sigma}\Delta_j + 2\mathbf{A}),\end{aligned}$$

where $\Delta_j = \tau_j - \tau_{j-1}$. This inspires us to consider the following state space model for the latent process $\mathbf{X}_t$ and observed process $\mathbf{Y}_t$ as follows:

$$\mathbf{Y}_j = \mathbf{X}_j + \mathbf{U}_j, \quad \mathbf{X}_j = \mathbf{X}_{j-1} + \mathbf{V}_j, \tag{2}$$

where $\mathbf{V}_j$s are independent, and normally distributed with mean zero and covariance $\mathbf{\Sigma}\Delta_j$. Here for simplicity in notations, we suppress the time τ by treating $\mathbf{Y}_{\tau_j} = (Y_{1\tau_j},\ldots,Y_{d\tau_j})'$ as $\mathbf{Y}_j = (Y_{1j},\ldots,Y_{dj})'$when no confusion arises, and the same convention applies for $\mathbf{X}_{\tau_j}$ and $\mathbf{U}_{\tau_j}$.

When applying model (2) to observed high frequency data, we treat the components with no observation at time τ_j as missing, and assume that there are d_j $(d_j \ge 1)$ observations at time τ_j for $j = 1,\ldots,n$. Let $\mathbf{B}_j = \binom{\mathbf{B}_j^{(1)}}{\mathbf{B}_j^{(2)}}$ be a permutation matrix such that $\tilde{\mathbf{Y}}_j = \mathbf{B}_j(Y_{1j}, Y_{2j},\ldots,Y_{dj})' = (\mathbf{Y}_j^{(1)'}, \mathbf{Y}_j^{(2)'})'$ where $\mathbf{Y}_j^{(1)} \in \mathbb{R}^{d_1}$ collects the observed asset prices, and $\mathbf{Y}_j^{(2)} \in \mathbb{R}^{d-d_j}$ is the missing component. Hence the role of $\mathbf{B}_j$ is such that the first d_j components in $\tilde{\mathbf{Y}}_j$ are observed. Then the state space model (2) can be written as

$$\tilde{\mathbf{Y}}_j = \mathbf{B}_j\mathbf{X}_j + \tilde{\mathbf{U}}_j, \quad \mathbf{X}_j = \mathbf{X}_{j-1} + \mathbf{V}_j, \tag{3}$$

where $\tilde{\mathbf{U}}_j$ is the reordered d dimensional random vector with mean zero and variance $\mathbf{A}_j = \mathbf{B}_j\mathbf{A}\mathbf{B}_j'$.

If all $\tilde{\mathbf{Y}}_j$ and $\mathbf{X}_j$ are observable, it is clear that under the assumption that the initial state $\mathbf{X}_0 \sim N(\boldsymbol{\mu}_*, \mathbf{\Sigma}_*)$ and ignore constant part, the log-likelihood function is given by

$$\begin{aligned}-2\ln L(\boldsymbol{\theta}) = {}& \ln|\mathbf{\Sigma}_*| + (\mathbf{X}_0 - \boldsymbol{\mu}_*)'\mathbf{\Sigma}_*^{-1}(\mathbf{X}_0 - \boldsymbol{\mu}_*) + n\ln|\mathbf{\Sigma}|\\ &+ \sum_{j=1}^{n} \Delta_j^{-1}(\mathbf{X}_j - \mathbf{X}_{j-1})'\mathbf{\Sigma}^{-1}(\mathbf{X}_j - \mathbf{X}_{j-1})\\ &+ n\ln|\mathbf{A}| + \sum_{j=1}^{n}\Big(\{\mathbf{B}_j'(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)\}'\mathbf{A}^{-1}\\ &\times\{\mathbf{B}_j'(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)\}\Big).\end{aligned} \tag{4}$$

where we denote by $\boldsymbol{\theta}$ the vector of parameters containing all elements of $\{\boldsymbol{\mu}_*, \mathbf{\Sigma}_*, \mathbf{\Sigma}, \mathbf{A}\}$. Then by taking the derivatives of parameters, we have the maximum likelihood estimators are $\hat{\mathbf{A}} = n^{-1}\sum_{j=1}^{n}\{\mathbf{B}_j'(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)\}\{\mathbf{B}_j'(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)\}'$ and $\hat{\mathbf{\Sigma}} = n^{-1}\sum_{j=1}^{n}\Delta_j^{-1}(\mathbf{X}_j - \mathbf{X}_{j-1})(\mathbf{X}_j - \mathbf{X}_{j-1})'$ respectively.

However, in practice one can only observe data $y_n = \{\mathbf{Y}_1^{(1)},\ldots,\mathbf{Y}_n^{(1)}\}$. To estimate the parameters, we apply the EM algorithm (Dempster, Laird, and Rubin, 1977). We first specify an initial value to $\boldsymbol{\theta}$ denoted by $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\mu}_*^{(0)}, \mathbf{\Sigma}_*^{(0)}, \mathbf{\Sigma}^{(0)}, \mathbf{A}^{(0)})$, where the ith element of $\boldsymbol{\mu}_*^{(0)}$ can simply be specified as the first observation of ith asset, $\mathbf{A}^{(0)}$ can be taken as the realized covariance divided by $2n$ (Zhang, Mykland and Aït-Sahalia, 2005), $\mathbf{\Sigma}_*$ and $\mathbf{\Sigma}$ can be initialized by the subsample based realized covariance matrix or the QML estimator (Aït-Sahalia, Fan, and Xiu, 2010; Liu and Tang, 2012). Let

$$\begin{aligned}\mathbf{X}_j^l &= E(\mathbf{X}_j|y_l) \quad \text{and}\\ \mathbf{P}_{j,m}^l &= E\{(\mathbf{X}_j - \mathbf{X}_j^l)(\mathbf{X}_m - \mathbf{X}_m^l)|y_l\},\end{aligned} \tag{5}$$

be the conditional expection and covariance given the data $y_l = \{\mathbf{Y}_1^{(1)},\ldots,\mathbf{Y}_l^{(1)}\}$, and denote by $\mathbf{P}_j^l = \mathbf{P}_{j,m}^l$ when $m = j$ the conditional variance.

The conditional exception of the log-likelihood function (4) given y_n and $\boldsymbol{\theta}^{(k-1)}$ requires evaluating the conditional expectations of $(\mathbf{X}_j - \mathbf{X}_{j-1})(\mathbf{X}_j - \mathbf{X}_{j-1})'$ and

$(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)'$. Specifically,

$$\begin{aligned}
&E\{(\mathbf{X}_j - \mathbf{X}_{j-1})(\mathbf{X}_j - \mathbf{X}_{j-1})'|y_n\} \\
&= E\left(\mathbf{X}_j\mathbf{X}_j' - \mathbf{X}_j\mathbf{X}_{j-1}' - \mathbf{X}_{j-1}\mathbf{X}_j' + \mathbf{X}_{j-1}\mathbf{X}_{j-1}'|y_n\right) \\
&= \left(\mathbf{X}_j^n\mathbf{X}_j^{n\prime} + \mathbf{P}_j^n\right) - \left(\mathbf{X}_j^n\mathbf{X}_{j-1}^{n}{}' + \mathbf{P}_{j,j-1}^n\right) \\
&\quad - \left(\mathbf{X}_{j-1}^n\mathbf{X}_j^{n\prime} + \mathbf{P}_{j-1,j}^n\right) + \left(\mathbf{X}_{j-1}^n\mathbf{X}_{j-1}^{n}{}' + \mathbf{P}_{j-1}^n\right).
\end{aligned}$$

Evaluating

$$E(\{\mathbf{B}_j'(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)\}\{\mathbf{B}_j'(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)\}'|y_n, \boldsymbol{\theta}^{(k-1)})$$

is not more involved since $\tilde{\mathbf{Y}}_j$ contains two parts—the observed components and missing components. By (3), we have

$$\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j = \begin{pmatrix} \mathbf{Y}_j^{(1)} - \mathbf{B}_j^{(1)}\mathbf{X}_j \\ \mathbf{Y}_j^{(2)} - \mathbf{B}_j^{(2)}\mathbf{X}_j \end{pmatrix} \sim N(0, \mathbf{A}_j).$$

By decomposing $\mathbf{A}_j$ into blocks according to the components in $\tilde{\mathbf{Y}}_j$ as $\left(\begin{smallmatrix}\mathbf{A}_{11j} & \mathbf{A}_{12j}\\ \mathbf{A}_{21j} & \mathbf{A}_{22j}\end{smallmatrix}\right)$, we have $\mathbf{Y}_j^{(2)}|\mathbf{X}_j$, $\mathbf{Y}_j^{(1)}$ follows normal distribution with mean $\mathbf{B}_j^{(2)}\mathbf{X}_j + \mathbf{A}_{21j}(\mathbf{A}_{11j})^{-1} \times (\mathbf{Y}_j^{(1)} - \mathbf{B}_j^{(1)}\mathbf{X}_j)$ and variance $\mathbf{A}_{22j} - \mathbf{A}_{21j}(\mathbf{A}_{11j})^{-1}\mathbf{A}_{12j}$.

Hence

$$\begin{aligned}
&E\left(\mathbf{Y}_j^{(2)} - \mathbf{B}_j^{(2)}\mathbf{X}_j|y_n\right) \\
&= E\left\{E\left(\mathbf{Y}_j^{(2)} - \mathbf{B}_j^{(2)}\mathbf{X}_j|y_n, \mathbf{X}_j = \mathbf{X}_j^n\right)|\mathbf{X}_j = \mathbf{X}_j^n\right\} \\
&= \mathbf{A}_{21j}(\mathbf{A}_{11j})^{-1}\left(\mathbf{Y}_j^{(1)} - \mathbf{B}_j^{(1)}\mathbf{X}_j^{(n)}\right).
\end{aligned}$$

And by the law of total variance, we have that

$$\begin{aligned}
&\mathrm{Var}\left(\mathbf{Y}_j^{(2)} - \mathbf{B}_j^{(2)}\mathbf{X}_j|y_n\right) \\
&= \mathrm{Var}\left\{E\left(\mathbf{Y}_j^{(2)} - \mathbf{B}_j^{(2)}\mathbf{X}_j|y_n, \mathbf{X}_j\right)\right\} \\
&\quad + E\left\{\mathrm{Var}\left(\mathbf{Y}_j^{(2)} - \mathbf{B}_j^{(2)}\mathbf{X}_j|y_n, \mathbf{X}_j\right)\right\}.
\end{aligned}$$

Therefore,

(6)

$$\begin{aligned}
&E_{k-1}\left\{(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)'|y_n\right\} \\
&= \left\{E_{k-1}\left(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j|y_n\right)E_{k-1}\left(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j|y_n\right)'\right\} \\
&\quad + \mathrm{Var}_{k-1}\left(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j|y_n\right) \\
&= \begin{pmatrix} \mathbf{Y}_j^{(1)} - \mathbf{B}_j^{(1)}\mathbf{X}_j^n \\ \mathbf{A}_{21j}^{(k-1)}\left(\mathbf{A}_{11j}^{(k-1)}\right)^{-1}\left(\mathbf{Y}_j^{(1)} - \mathbf{B}_j^{(1)}\mathbf{X}_j^n\right) \end{pmatrix} \\
&\quad \times \begin{pmatrix} \mathbf{Y}_j^{(1)} - \mathbf{B}_j^{(1)}\mathbf{X}_j^n \\ \mathbf{A}_{21j}^{(k-1)}\left(\mathbf{A}_{11j}^{(k-1)}\right)^{-1}\left(\mathbf{Y}_j^{(1)} - \mathbf{B}_j^{(1)}\mathbf{X}_j^n\right) \end{pmatrix}' \\
&\quad + \begin{pmatrix} \mathbf{B}_j^{(1)} \\ \mathbf{A}_{21j}^{k-1}\left(\mathbf{A}_{11j}^{(k-1)}\right)^{-1}\mathbf{B}_j^{(1)} \end{pmatrix}\mathbf{P}_j^n\begin{pmatrix} \mathbf{B}_j^{(1)} \\ \mathbf{A}_{21j}^{k-1}\left(\mathbf{A}_{11j}^{(k-1)}\right)^{-1}\mathbf{B}_j^{(1)} \end{pmatrix}' \\
&\quad + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22j}^{(k-1)} - \mathbf{A}_{21j}^{(k-1)}\left(\mathbf{A}_{11j}^{(k-1)}\right)^{-1}\mathbf{A}_{12j}^{(k-1)} \end{pmatrix},
\end{aligned}$$

where E_{k-1} and Var_{k-1} means the expectation and variance are taken under $\boldsymbol{\theta}^{(k-1)}$.

Let $\mathbf{M}^{(k)} = E_{k-1}\{(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)'|y_n\}$. Then the E-step of EM algorithm at kth iteration is

(7)
$$\begin{aligned}
&Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)}) \\
&= E\left(-2\ln L_x(\boldsymbol{\theta})|y_n, \boldsymbol{\theta}^{(k-1)}\right) \\
&= \ln|\boldsymbol{\Sigma}_0| + \mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}\{\mathbf{P}_0^n + (\mathbf{X}_0^n - \boldsymbol{\mu}_0)(\mathbf{X}_0^n - \boldsymbol{\mu}_0)'\}\right) \\
&\quad + n\ln|\boldsymbol{\Sigma}| + \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{S}_{11} - \mathbf{S}_{10} - \mathbf{S}_{10}' + \mathbf{S}_{00})\right\} \\
&\quad + n\ln|\mathbf{A}| + \mathrm{tr}\left(\mathbf{A}^{-1}\mathbf{B}_j'\mathbf{M}^{(k)}\mathbf{B}_j\right),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{S}_{11} &= \sum_{j=1}^n \Delta_j^{-1}(\mathbf{P}_j^n + \mathbf{X}_j^n\mathbf{X}_j^{n\prime}), \\
\mathbf{S}_{00} &= \sum_{j=1}^n \Delta_j^{-1}(\mathbf{P}_{j-1}^n + \mathbf{X}_{j-1}^n\mathbf{X}_{j-1}^{n}{}'), \\
\mathbf{S}_{10} &= \sum_{j=1}^n \Delta_j^{-1}(\mathbf{X}_j^n\mathbf{X}_{j-1}^{n}{}' + \mathbf{P}_{j,j-1}^n).
\end{aligned}$$

Then by solving the first order condition, we complete the M-step and update the estimators of $\boldsymbol{\Sigma}$ and $\mathbf{A}$ by

(8)
$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}^{(k)} &= n^{-1}(\mathbf{S}_{11} + \mathbf{S}_{00} - \mathbf{S}_{10} - \mathbf{S}_{10}') \\
\hat{\mathbf{A}}^{(k)} &= n^{-1}\sum_{j=1}^n \mathbf{B}_j'\mathbf{M}^{(k)}\mathbf{B}_j,
\end{aligned}$$

and update $\boldsymbol{\mu}_*$ and $\boldsymbol{\Sigma}_*$ by

(9)
$$\hat{\boldsymbol{\mu}}_*^{(k)} = \mathbf{X}_0^n \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_*^{(k)} = \mathbf{P}_0^n.$$

Then the final estimator for $\boldsymbol{\Sigma}$ and $\mathbf{A}$ are obtained by repeating the E-step and M-step until convergence.

We note that the conditional expectations in (5) can be conveniently evaluated by the Kalman filtering method; see, for example, Shumway and Stoffer (2006). Since our state space model is constructed from a quasi-maximum likelihood approach, we call it the QKF approach. For completeness of the framework, we present the results for applying the Kalman filtering in the Appendix.

We make the following remarks on the state space model approach.

Remark 1. When $\mathbf{A}$ is a diagonal matrix, i.e. components in $\mathbf{U}_t$ are uncorrelated. The E-step is simplified by observing that

$$\begin{aligned}\mathbf{M}^{(k)} &= E_{k-1}\left\{(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)(\tilde{\mathbf{Y}}_j - \mathbf{B}_j\mathbf{X}_j)'|y_n\right\} \\ &= (\tilde{\mathbf{Y}}_j - \tilde{\mathbf{B}}_j\mathbf{X}_j^n)(\tilde{\mathbf{Y}}_j - \tilde{\mathbf{B}}_j\mathbf{X}_j^n)' + \tilde{\mathbf{B}}_j\tilde{\mathbf{P}}_j^n\tilde{\mathbf{B}}_j' \\ &\quad + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22j}^{(k-1)} \end{pmatrix},\end{aligned}$$

with $\tilde{\mathbf{Y}}_j = \binom{\mathbf{Y}_j^{(1)}}{\mathbf{0}}$ and $\tilde{\mathbf{B}}_j = \binom{\mathbf{B}_j^{(1)}}{\mathbf{0}}$. This is actually the case considered in Shepard and Xiu (2012) and Crosi, Peluso, and Audrino (2012). A closer look at (6) reveals that when $\mathbf{A}$ is not a diagonal matrix, information in the observed components $\mathbf{Y}_j^{(1)}$ is incorporated when taking the conditional expectation of $\mathbf{Y}_j^{(2)}$. This fact may provide an opportunity for efficiency gain. In addition, if the matrix $\mathbf{A}$ is misspecified as a diagonal matrix, it is likely a bias may incur in the estimator $\hat{\mathbf{\Sigma}}$; see our simulation studies for more detail.

Remark 2. Theory of the EM algorithm (Wu, 1983) ensures the convergence of the estimator to the maximizer of the so-called complete data likelihood function which is constructed based on y_n. Thus, the estimator based on the state space model approach is consistent as long as the complete data likelihood based estimator is consistent. To our best knowledge, there is no general theory ensuring the consistency of the estimators using the EM algorithm. Most likely this is because the theoretical analysis of the estimator with missing data depends on the specific mechanism of the data missingness. Nevertheless, in the high frequency financial data case, data missingness from how transactions are triggered is an interesting and challenging problem; see, for example, Engle and Russell (1998). On the other hand, as shown in Liu and Tang (2012), the maximum likelihood approach with appropriately synchronized data is consistent under some conditions on the data asynchronicity. Since a data synchronization scheme only incorporates a portion of the data information contained in the complete data, we may reasonably conjecture that the estimator based on the state space model approach is also consistent because it incorporates more data information.

Remark 3. A family of methods with multivariate high frequency financial data is applying data pre-precessing with synchronization; see, for example, Barndorff-Nielsen, Hansen, Lunde, and Shepard (2011), Zhang (2011), Liu and Tang (2012). Inevitably, a portion of data is deleted and thus information loss may incur. As shown in our simulation studies, the efficiency gain is substantial by using the state space model approach with EM algorithm as compared with the approach using data synchronization.

Remark 4. When all components in $\mathbf{Y}$ are observed with equally spaced data, it can be shown that the state space model approach is equivalent to the quasi-maximum likelihood approach in Liu and Tang (2012). Let

$$\overline{\mathbf{Y}}_j = \mathbf{Y}_j - \mathbf{Y}_{j-1} = \mathbf{X}_j - \mathbf{X}_{j-1} + \mathbf{U}_j - \mathbf{U}_{j-1} \ (j = 1, \ldots, n)$$

be the log-returns of d assets and $\overline{\mathbf{Y}} = (\overline{\mathbf{Y}}_1', \overline{\mathbf{Y}}_2', \ldots, \overline{\mathbf{Y}}_n')'$. Then the log-likelihood function of $\overline{\mathbf{Y}}$ is

$$-2\ln L_{\overline{\mathbf{Y}}}(\mathbf{\Sigma}, \mathbf{A}) = \ln|\mathbf{\Omega}| + \overline{\mathbf{Y}}'\mathbf{\Omega}^{-1}\overline{\mathbf{Y}}, \tag{10}$$

where

$$\mathbf{\Omega} = \mathbf{I}_n \otimes (\mathbf{\Sigma}\Delta + 2\mathbf{A}) - (\mathbf{L}_n + \mathbf{L}_n') \otimes \mathbf{A},$$

where $\otimes$ denotes the Kronecker product, $\mathbf{I}_n$ is the n-dimensional identity matrix, $\mathbf{L}_n = (L_{kl})$ $(k, l = 1, \ldots, n)$ is an n-dimensional one-lag sub-diagonal matrix with $L_{k-1,k} = 1$ $(k = 2, \ldots, n)$ and all other elements being 0. Hence, by results in Liu and Tang (2012), the state space model approach for synchronous data is also consistent in this special case, and also achieves the optimal rate of convergence $n^{1/4}$ in the sense of Gloter and Jacod (2001).

Remark 5. The estimators of $\mathbf{\Sigma}$ and $\mathbf{A}$ are positive semi-definite by observing that in the explicit forms (8), the updates in each M-step are all positive semi-definite.

Remark 6. In developing the EM algorithm for estimating the ICM, it is assumed that the true log returns and the microstructure noise respectively follow normal distributions. In literature, the normal distribution is conventionally used in studying financial returns (Black and Scholes, 1973) mainly because of its convenience for more tractable analysis. In our study, the main role of the normal assumption is to ensure explicit forms in the EM algorithm by using Kalman filtering. We note that the validity of the QML approach for estimating the ICM does not require the distributional assumption to be true; see also Xiu (2010) and Liu and Tang (2012). As shown in our simulations, the state space model approach works very well even when returns do not follow normal distributions.

3. NUMERICAL EXAMPLES

3.1 Simulations

We demonstrate the mertis of the state space model approach (QKF) by extensive simulations. We compare the state space model approach to the quasi-maximum likelihood approach (QML) of Liu and Tang (2012), and the approach of Aït-Sahalia, Fan, and Xiu (2010) that utilizes a polarization identity and so is denoted by POL. The number of replications in simulations is 1,000 for call cases. The QML approach estimate $\mathbf{\Sigma}$ and $\mathbf{A}$ by maximizing the quasi-log-likelihod function (10); see Liu and Tang (2012) for efficient algorithm for numerical implementations. The POL approach estimates the ICM element by element. Specifically, the integrated covariation of latent processes X_{kt} and

Table 1. *Values for the model parameters in the simulations*

Asset	κ_i	s_i	$\bar{\sigma}_i^2$	λ_i	μ_i	θ_i	a_i	ρ_i
1	6	0.5	0.25	12	0.8	−5	0.005	−0.3
2	4	0.3	0.16	36	1.2	−6	0.003	−0.2
3	5	0.4	0.09	24	0.1	−7	0.004	−0.15

X_{lt} in the two asset log-price processes Y_{kt} and Y_{lt}, and the covariance of noises U_{kt} and U_{lt} contaminated in the observations of Y_{kt} and Y_{lt} are estimated by using the polarization identity for random variables:

$$\widehat{\text{cov}}(Z_k, Z_l) = \{\widehat{\text{var}}(\gamma Z_k + (1-\gamma)Z_l) + \widehat{\text{var}}(\gamma Z_k - (1-\gamma)Z_l)\} / \{4\gamma(1-\gamma)\}$$

where γ can be chosen as $\widehat{\text{var}}(Z_l)/\{\widehat{\text{var}}(Z_k) + \widehat{\text{var}}(Z_l)\}$ or other values in practice, $\widehat{\text{var}}(\gamma Z_k \pm (1-\gamma)Z_l)$ is the one dimensional QML estimator (Aït-Sahalia, Mykland, and Zhang, 2005; Xiu, 2010) of the integrated variances of latent process in the new series $\gamma Z_{kt} \pm (1-\gamma)Z_{lt}$ respectively. We note that both QML approach and the POL approach require synchronized data so that some data synchronization scheme is required if handing experiments with asynchronous data.

We consider in the first experiment the performances of three estimators in estimating the ICM for equally spaced synchronous data. We generate data of the log-price process from the Heston Model,

$$dX_{it} = \sigma_{it} dW_{it}, \ (i = 1, 2, 3)$$
$$d\sigma_{it}^2 = \kappa_i(\bar{\sigma}_i^2 - \sigma_{it}^2)dt + s_i \sigma_{it} dB_{it} + \sigma_{it-} J_{it}^V dN_{it},$$

where $E(dW_{it} \cdot dB_{jt}) = \delta_{ij}\rho_i dt$, $\delta_{ij} = 1$ for $i = j$; $\delta_{ij} = 0$ for $i \neq j$ and $E(dW_{it} \cdot dW_{jt}) = \rho_{ij} dt$. The first observation of volatility process σ_{i0}^2 is sampled from a Gamma distribution $\Gamma(\kappa_i \bar{\sigma}_i^2 / s_i^2, s_i^2 / 2\kappa_i)$. The jump size J_{it}^V in volatility equals to $\exp(z_i)$, where $z_i \sim N(\theta_i, \mu_i)$, and N_{it} is a Poisson Process independent of other processes with intensity λ_i. The parameters are respectively specified by values in Table 1. The noises $\{U_t\}_{t=1}^n$ are independent and identically distributed with distribution $N(0, \mathbf{A})$, where $A_{ij} = a_i a_j \breve{\rho}_{ij}$ with a_i for $i = 1, 2, 3$ are given in Table 1 and $\breve{\rho}_{12} = -0.2$, $\breve{\rho}_{13} = -0.15$ and $\breve{\rho}_{23} = 0.1$. We calculate the bias and root mean square error (RMSE) for the three approaches, and also compute the relative efficiency (RE) to compare three approaches where REQ is the ratio of the RMSE of QML and QKF and REP is the ration of RMSE of POL and QKF. Therefore a relative efficiency with value greater than 1 indicates a better performance of the QKF approach. We consider $d = 2$ by only using the first two log-price processes generated, and $d = 3$ by considering all three processes. We vary the correlation between two latent log-return processes to compare the performances of estimators when $d = 2$. When $d = 3$, we set the correlations as $\rho_{12} = 0.3, \rho_{13} = 0.6, \rho_{23} = 0.9$. Results for the two dimensional case are reported in Table 2, and results for the three dimensional case are reported in the upper part of Table 3. We can see that those results are consistent with our expectations. First, the QML and QKF approach perform similarly while both have better performance than the POL approach especially when the correlation level between processes is higher and the sampling interval is smaller. This demonstrates the advantage of the approaches that utilizes information from the quasi-likelihood function. In this experiment, we also note that the QKF and QML approaches have close performance to each other, which is also expected.

We conduct the second experiment to assess the performance of the methods with asynchronous data. We firstly generate original equally spaced synchronous log-prices data

Table 2. *Biases and root mean square errors (RMSE, values in brackets) ($\times 10^2$) of the bivariate cases for elements of ICM $[\hat{\Sigma}_{11}, \hat{\Sigma}_{12}, \hat{\Sigma}_{22}]$ when data are synchronous and equally spaced with time interval between two consecutive observations equals to Δ and correlation between two log-price processes equals to ρ*

Syn	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{12}$	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{12}$
	$\Delta = 2s$			$\Delta = 12s$		
	$\rho = 0.3$			$\rho = 0.3$		
POL	0.03	0.03	0.05	0.01	0.01	0.01
	(0.80)	(0.51)	(0.46)	(1.65)	(1.04)	(0.97)
QLE	0.02	0.03	0.06	0.01	0.01	0.01
	(0.76)	(0.49)	(0.37)	(1.58)	(0.98)	(0.62)
QKF	0.01	0.04	0.04	0.01	0.01	0.01
	(0.75)	(0.50)	(0.35)	(1.57)	(0.93)	(0.60)
REP	1.05	1.04	1.20	1.05	1.12	1.62
REQ	1.01	0.98	1.06	1.01	1.05	1.03
	$\rho = 0.6$			$\rho = 0.6$		
POL	0.02	0.00	0.03	0.04	0.05	0.05
	(0.84)	(0.48)	(0.50)	(1.62)	(1.03)	(1.01)
QML	0.02	0.00	0.02	0.04	0.04	0.04
	(0.74)	(0.43)	(0.41)	(1.39)	(0.87)	(0.66)
QKF	0.02	0.01	0.03	0.03	0.05	0.05
	(0.75)	(0.44)	(0.40)	(1.40)	(0.88)	(0.65)
REP	1.12	1.11	1.18	1.16	1.17	1.55
REQ	0.99	0.98	1.03	0.99	0.99	1.02
	$\rho = 0.9$			$\rho = 0.9$		
POL	0.03	0.03	0.02	0.01	0.00	0.00
	(0.83)	(0.50)	(0.57)	(1.63)	(1.04)	(1.17)
QML	0.02	0.02	0.02	0.04	0.02	0.03
	(0.66)	(0.40)	(0.44)	(1.10)	(0.74)	(0.73)
QKF	0.01	0.01	0.03	0.03	0.02	0.05
	(0.67)	(0.39)	(0.45)	(1.13)	(0.74)	(0.73)
REP	1.21	1.20	1.23	1.44	1.41	1.60
REQ	0.99	1.03	0.95	0.97	1.00	1.00

Note: The left 3 columns and right 3 columns are results obtained by using data which is generated with time interval $\Delta = 2s$ and $\Delta = 12s$ respectively. QKF is our new estimator obtained by combining the QML approach and Kalman filter together. QML is the estimator developed in Liu and Tang (2012) and POL is the estimator derived in Aït-Sahalia, Fan, and Xiu (2010). REP is the ratio of the RMSEs of POL and RMSE of QKF and REQ is the ratio of the RMSEs of QLE and RMSE of QKF.

Table 3. Bias and root mean square errors (RMSE, values in brackets) ($\times 10^2$) of the three processes case for elements of ICM $[\hat{\Sigma}_{ij}]$ $(i,j=1,2,3)$ when data are synchronous and asynchronous

	Synchronous Data					
$\Delta = 12s$	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{33}$	$\hat{\Sigma}_{12}$	$\hat{\Sigma}_{13}$	$\hat{\Sigma}_{23}$
POL	0.01	0.09	0.01	0.01	0.18	0.17
	(1.56)	(1.05)	(0.63)	(0.96)	(0.80)	(0.71)
QML	0.01	0.09	0.02	0.18	0.17	0.01
	(1.47)	(0.93)	(0.49)	(0.84)	(0.69)	(0.59)
QKF	0.01	0.03	0.02	0.12	0.25	0.25
	(1.46)	(0.90)	(0.52)	(0.84)	(0.70)	(0.60)
REP	1.07	1.17	1.21	1.14	1.14	1.18
REQ	1.01	1.03	0.95	1.00	0.98	0.98
	Asynchronous Data					
$\Delta = 6s$	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{33}$	$\hat{\Sigma}_{12}$	$\hat{\Sigma}_{13}$	$\hat{\Sigma}_{23}$
POL	0.12	0.01	0.07	0.15	0.29	0.28
	(1.80)	(1.15)	(0.75)	(1.10)	(0.87)	(0.80)
QML	0.20	0.08	0.09	0.21	0.34	0.31
	(1.72)	(1.10)	(0.71)	(1.02)	(0.81)	(0.77)
QKF	0.21	0.04	0.05	0.17	0.31	0.27
	(1.24)	(0.75)	(0.53)	(0.68)	(0.57)	(0.52)
REP	1.45	1.53	1.42	1.62	1.53	1.54
REQ	1.39	1.47	1.34	1.50	1.42	1.48

Note: The upper part are the results for equally spaced synchronous data with time interval between two consecutive observations equals to $\Delta = 12s$, and the bottom part are the results for asynchronous data generated through Bernoulli trials with success probabilities $0.6, 0.8, 0.5$ for three assets from original equally spaced synchronous data with time interval between two consecutive observations equals to $\Delta = 6s$. QKF is our new estimator obtained by combining the QML approach and Kalman filter together. QML is the estimator developed in Liu and Tang (2012) and POL is the estimator derived in Aït-Sahalia, Fan, and Xiu (2010). REP is the ratio of the RMSEs of POL and RMSE of QKF and REQ is the ratio of the RMSEs of QLE and RMSE of QKF.

for three assets processes by choosing the time interval $\Delta = 2s$ or $\Delta = 12s$ in the same way as that in the first case, and then we use Bernoulli trials with success probabilities p_1, p_2, p_3 to randomly select observations from original data for three log-price processes respectively. We then use refresh time scheme (Barndorff-Nielsen, Hansen, Lunde, and Shepard, 2011) to synchronize them for the QML and POL approaches. Table 4 reports the results comparing the three approaches under different correlation ρ_{12} and the same (p_1, p_2). Table 5 reports the results comparing the three approaches under different (p_1, p_2) and the same correlation ρ_{12}. The lower part of Table 3 reports the results of the three approaches for 3 dimensional ICM when data are asynchronous. This experiment confirms the promising performance of the QKF approach especially when $d = 3$ (Table 3) and data with higher level asynchronicity (Table 5). Therefore, we clearly see the improvement of the state space model approach without requiring data synchronization so that data information is most efficiently incorporated.

Table 4. Bias and root mean square errors (RMSE, values in brackets) ($\times 10^2$) of the two processes case for elements of ICM $[\hat{\Sigma}_{11}, \hat{\Sigma}_{12}, \hat{\Sigma}_{22}]$ when data are asynchronous

Asyn	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{12}$	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{12}$	$\hat{\Sigma}_{22}$
	$\Delta = 2s$			$\Delta = 12s$		
	$\rho = 0.3$			$\rho = 0.3$		
POL	0.08	0.08	0.00	0.08	0.01	0.01
	(1.13)	(0.69)	(0.62)	(2.21)	(1.40)	(1.27)
QLE	0.02	0.03	0.06	0.01	0.01	0.01
	(1.10)	(0.68)	(0.48)	(2.19)	(1.35)	(1.20)
QKF	0.04	0.03	0.07	0.01	0.01	0.01
	(0.99)	(0.54)	(0.45)	(1.90)	(1.01)	(0.85)
REP	1.14	1.28	1.38	1.16	1.39	1.49
REQ	1.11	1.26	1.07	1.15	1.34	1.41
	$\rho = 0.6$			$\rho = 0.6$		
POL	0.02	0.00	0.03	0.04	0.05	0.05
	(1.15)	(0.72)	(0.71)	(2.16)	(1.43)	(1.35)
QML	0.02	0.00	0.02	0.04	0.04	0.04
	(1.03)	(0.65)	(0.53)	(2.29)	(1.35)	(1.28)
QKF	0.02	0.00	0.03	0.04	0.05	0.05
	(0.94)	(0.52)	(0.48)	(1.76)	(0.96)	(0.87)
REP	1.22	1.38	1.48	1.23	1.49	1.55
REQ	1.10	1.25	1.10	1.30	1.41	1.47
	$\rho = 0.9$			$\rho = 0.9$		
POL	0.03	0.03	0.02	0.01	0.00	0.00
	(1.10)	(0.70)	(0.77)	(2.16)	(1.35)	(1.44)
QML	0.02	0.02	0.02	0.04	0.02	0.03
	(0.87)	(0.56)	(0.57)	(2.00)	(1.20)	(1.30)
QKF	0.02	0.02	0.02	0.04	0.02	0.03
	(0.78)	(0.45)	(0.52)	(1.42)	(0.87)	(0.89)
REP	1.41	1.56	1.48	1.52	1.55	1.62
REQ	1.12	1.24	1.10	1.41	1.38	1.46

Note: The left 3 columns and right 3 columns are results obtained by using the asynchronous data generated through Bernoulli trials with success probabilities $p_1 = 0.5$ for the first asset and $p_2 = 0.8$ for the second asset from original data, which are synchronous and equally spaced with time interval between two consecutive observations equals to $\Delta = 2s$ and $\Delta = 12s$ respectively, and the correlation between two log-price processes equals to ρ. QKF is our new estimator obtained by combining the QML approach and Kalman filter together. QML is the estimator developed in Liu and Tang (2012) and POL is the estimator derived in Aït-Sahalia, Fan, and Xiu (2010). REP is the ratio of the RMSEs of POL and RMSE of QKF and REQ is the ratio of the RMSEs of QLE and RMSE of QKF.

In the third experiment, we investigate the impact of the specification of the covariance matrix $\mathbf{A}$ of the microstructure noises for the QKF and QML approaches. For such a purpose, we generate data with a non-diagonal matrix $\mathbf{A}_0$, while in the estimation methods $\mathbf{A}$ is specified as diagonal. In this experiment, data are generated in the same way as that in the second experiment, except that the correlation between the noises in observations of two assets is specified as $\breve{\rho} = -0.5$ and the correlation between the two latent log-returns process is fixed as $\rho = 0.6$. Results for this experiment are reported in Table 6. The remarkable finding is that if $\mathbf{A}$ is misspecified, the performances of the QML and QKF

Table 5. Bias and root mean square errors (RMSE, values in brackets) ($\times 10^2$) of the two processes case for elements of ICM $[\hat{\Sigma}_{11}, \hat{\Sigma}_{12}, \hat{\Sigma}_{22}]$ when data are asynchronous

Asyn	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{12}$	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{12}$
		$\Delta = 2s$			$\rho = 0.6$	
		$p_1 = 0.5$			$p_2 = 0.8$	
POL	0.02	0.00	0.03	0.04	0.05	0.05
	(1.30)	(0.78)	(0.67)	(2.16)	(1.43)	(1.35)
QML	0.02	0.00	0.02	0.04	0.04	0.04
	(1.25)	(0.79)	(0.66)	(2.29)	(1.35)	(1.28)
QKF	0.02	0.00	0.03	0.04	0.05	0.05
	(0.73)	(0.33)	(0.26)	(3.94)	(1.58)	(1.51)
REP	1.22	1.38	1.48	1.23	1.49	1.55
REQ	1.10	1.25	1.10	1.30	1.41	1.47
		$p_1 = 0.8$			$p_2 = 0.3$	
POL	0.03	0.03	0.02	0.01	0.00	0.00
	(1.39)	(0.86)	(0.88)	(2.86)	(1.68)	(1.73)
QML	0.02	0.02	0.02	0.04	0.02	0.03
	(1.37)	(0.86)	(0.87)	(2.74)	(1.63)	(1.68)
QKF	0.02	0.02	0.02	0.04	0.02	0.03
QKF	(0.86)	(0.56)	(0.60)	(1.62)	(1.09)	(1.23)
REP	1.62	1.54	1.47	1.77	1.54	1.41
REQ	1.59	1.54	1.45	1.69	1.50	1.37

Note: The left 3 columns and right 3 columns are results obtained by using the asynchronous data generated through Bernoulli trials with success probabilities p_1 for the first asset and p_2 for the second asset from original data, which are synchronous and equally spaced with time interval between two consecutive observations equals to $\Delta = 2s$ and $\Delta = 12s$ respectively, and the correlation between two log-price processes equals to ρ.

Table 6. Ratios ($\times 100$) of biases and true values, and ratios ($\times 100$) of root mean square errors and true values for elements of ICM $[\hat{\Sigma}_{11}, \hat{\Sigma}_{12}, \hat{\Sigma}_{22}]$ when data are asynchronous

Parameters		$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{12}$
$\Delta = 12s$	POL	0.23(9.30)	0.52(8.66)	3.58(10.4)
$\breve{\rho} = -0.5$	QML	7.40(11.1)	6.98(10.4)	15.8(17.7)
$p_1 = 0.5$	QKF_1	5.31(8.60)	3.14(5.99)	12.3(14.0)
$p_2 = 0.8$	QKF_2	0.62(6.16)	0.58(4.71)	1.75(7.25)
$\Delta = 4s$	POL	0.19(5.74)	0.00(5.64)	1.51(7.48)
$\breve{\rho} = -0.5$	QML	10.6(11.7)	9.83(11.0)	25.0(25.6)
$p_1 = 0.5$	QKF_1	8.57(9.50)	6.05(6.92)	22.2(22.7)
$p_2 = 0.8$	QKF_2	0.23(3.84)	0.17(3.08)	1.51(4.62)

Note: Results are 100 times the actual values, and obtained by using the asynchronous data generated through Bernoulli trials with success probabilities p_1 and p_2 for the first and second assets respectively from original data, which are synchronous and equally spaced with time interval between two consecutive observations equals to Δ, $\breve{\rho}$ is the correlation between the noise in two assets. POL, QML, and QKF_1 are obtained by assuming the covariance matrix of noises is a diagonal matrix, and QKF_2 is obtained by assuming the covariance matrix of noises is a general covariance matrix.

approaches become worse when the sampling interval gets smaller. Since smaller sampling time interval means larger sample size, results in Table 6 clearly indicate a systematic bias due to the misspecification of $\mathbf{A}$. In contrast, we can see a better performance of the QKF approach that is with improved results with smaller Δ.

Table 7. Correlation matrix of ICM for 10 assets log-return process (values with stars at right head are minus)

$$\begin{pmatrix}
1 & 0.62^* & 0.45 & 0.34^* & 0.26 & 0.20^* & 0.16 & 0.13^* & 0.10 & 0.08^* \\
 & 1 & 0.72^* & 0.54 & 0.42^* & 0.33 & 0.26^* & 0.20 & 0.16^* & 0.13 \\
 & & 1 & 0.75^* & 0.58 & 0.46^* & 0.36 & 0.29^* & 0.23 & 0.18^* \\
 & & & 1 & 0.77^* & 0.61 & 0.48^* & 0.38 & 0.30^* & 0.24 \\
 & & & & 1 & 0.78^* & 0.62 & 0.49^* & 0.39 & 0.31^* \\
 & & & & & 1 & 0.79^* & 0.63 & 0.50^* & 0.40 \\
 & & & & & & 1 & 0.79^* & 0.63 & 0.50^* \\
 & & & & & & & 1 & 0.80^* & 0.63 \\
 & & & & & & & & 1 & 0.80^* \\
 & & & & & & & & & 1
\end{pmatrix}$$

Table 8. Ratios of root mean square errors of the POL approach and the QKF approach for elements of ICM when data are synchronous and equally spaced with time interval between two consecutive data equals to 12s

$$\begin{pmatrix}
1.09 & 1.13 & 1.15 & 1.20 & 1.21 & 1.20 & 1.19 & 1.19 & 1.18 & 1.17 \\
 & 1.16 & 1.21 & 1.24 & 1.26 & 1.25 & 1.25 & 1.24 & 1.22 & 1.22 \\
 & & 1.21 & 1.30 & 1.35 & 1.32 & 1.31 & 1.29 & 1.27 & 1.26 \\
 & & & 1.29 & 1.37 & 1.37 & 1.38 & 1.35 & 1.32 & 1.28 \\
 & & & & 1.37 & 1.39 & 1.42 & 1.41 & 1.38 & 1.33 \\
 & & & & & 1.33 & 1.37 & 1.37 & 1.36 & 1.32 \\
 & & & & & & 1.33 & 1.38 & 1.38 & 1.36 \\
 & & & & & & & 1.33 & 1.34 & 1.34 \\
 & & & & & & & & 1.31 & 1.33 \\
 & & & & & & & & & 1.26
\end{pmatrix}$$

In the last experiment, we conduct a simulation for $d = 10$. In this case, it is shown that the POL approach actually outperforms the QML approach (Liu and Tang, 2012) due to the large portion of data deleted in the data synchronization for the QML approach. Therefore, we only compare the performance between the POL approach and the QKF approach. We construct the 20×20 dimensional correlation matrix of $(d\mathbf{W}', d\mathbf{B}')'$ as $\begin{pmatrix} \mathbf{C}_W & \text{diag}(\boldsymbol{\rho}) \\ \text{diag}(\boldsymbol{\rho}) & \mathbf{I}_{10} \end{pmatrix}$, where $\mathbf{C}_W$ is given in Table 7, and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{10})'$ and other parameters in the Heston model are specified by setting β_j to be $(\sum_{i=1}^{3} \beta_i/3)(1 + |j-5|0.01)$ for the jth asset with β_i representing $\kappa_i, s_i, \bar{\sigma}_i^2, \lambda_i, \mu_i, \theta_i, a_i, \rho_i$ whose values are given in Table 1. We generate the synchronous data the same as in the first experiment and the asynchronous data the same as in the second experiment. Since the biases of the POL approach and the QKF approach are all small which are less than 5% of true values, we only report the relative efficiency (RE) of the QKF approach which is ratio of the RMSE of the QKF estimator and the POL estimator. Table 8 displays the REs for equally spaced synchronous data with time interval $\Delta = 12$ seconds. Table 9 are the relative efficiencies for asynchronous data which are generated through Bernoulli trials with success probability for the ith asset to be $0.6(1 + |i-5|0.02)$ $(i = 1, \ldots, 10)$ from original equally spaced synchronous data with time interval $\Delta = 6$ seconds. Refresh time scheme is then applied to synchronize the asynchronous data. We find that both Tables 8 and 9 confirm the

Table 9. Ratios of root mean square errors of the POL approach and the QKF approach for elements of ICM with asynchronous data generated through Bernoulli trials with success probability equal to $0.6(1+|i-5|0.02)$ *for the* i*th asset from original equally spaced synchronous data with time interval between two consecutive observations equals to 6s*

$$\begin{pmatrix} 1.44 & 1.48 & 1.51 & 1.56 & 1.55 & 1.55 & 1.60 & 1.64 & 1.66 & 1.56 \\ & 1.49 & 1.53 & 1.59 & 1.66 & 1.64 & 1.68 & 1.68 & 1.68 & 1.62 \\ & & 1.65 & 1.61 & 1.68 & 1.69 & 1.77 & 1.80 & 1.71 & 1.68 \\ & & & 1.75 & 1.63 & 1.66 & 1.81 & 1.87 & 1.84 & 1.81 \\ & & & & 1.93 & 1.64 & 1.82 & 1.86 & 1.82 & 1.87 \\ & & & & & 1.82 & 1.68 & 1.76 & 1.73 & 1.78 \\ & & & & & & 1.76 & 1.76 & 1.75 & 1.75 \\ & & & & & & & 1.81 & 1.69 & 1.73 \\ & & & & & & & & 1.85 & 1.74 \\ & & & & & & & & & 1.75 \end{pmatrix}$$

promising performance of the QKF approach that substantially improves the POL approach which requires minimal amount of data synchronization. This demonstrates the advantage and efficiency gain of the QKF approach without requiring data synchronization so that it utilizes all information for the observations.

3.2 Financial data analysis

We now illustrate the state space model approach in a financial trading data set with three stocks: IBM, Dell, and Microsoft. The data are obtained from the TAQ database where the first two trading days of 2007—January 4th and 5th—are considered. We organize three processes as $\mathbf{Y}=$ (IBM, Dell, Microsoft)$'$, so $\hat{\Sigma}_{11}$ is the estimator of integrated volatility of IBM, and so on, $\hat{\Sigma}_{ij}$ and $\hat{\rho}_{ij}$ $(i,j=1,2,3)$ are the estimates of the corresponding covariance and correlation. We conduct the same cleaning procedure as in Barndorff-Nielsen, Hansen, Lunde, and Shepard (2011) before applying the methods. For comparison purposes, we also implement the method of in Aït-Sahalia, Fan, and Xiu (2010) which is denoted by POL since it utilizes a polarization identity. We also compare the proposed approach to the quasi-maximum likelihood approach of Liu and Tang (2012) which is denoted by QML. We note that both the POL and QML approaches require data synchronization, but the former only needs synchronized pairs. The refresh time scheme of (Barndorff-Nielsen, Hansen, Lunde, and Shepard, 2011) is applied for pre-processing the data for approaches that require synchronized data. Results are reported in Table 10 where the estimates of the ICM are multiplied by 252 for annualizing the volatilities and covariations.

From Table 10, we can see the three approaches estimate the volatilities and covariations with different values. On both days, the state space model approach estimates the volatilities with larger values while the estimates on January 5 for DELL is substantially higher than methods using synchronized data. This is the remarkable difference between methods because portion of data is deleted in the synchronization procedure. In the original data set, the number of

Table 10. Estimators for the elements of ICM $[\hat{\Sigma}_{ii}]\times 252$ *and the correlations* $\breve{\rho}^*_{ij}$ $(i,j=1,2,3)$ *of IBM, DELL and MFT for empirical study in Section 3.2*

04/01/2007	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{33}$	$\hat{\rho}_{12}$	$\hat{\rho}_{13}$	$\hat{\rho}_{23}$
POL	3.03	7.04	3.26	0.24	0.22	0.37
QML_1	3.02	7.13	3.15	0.20	0.19	0.36
QML_2	3.00	7.04	3.10	0.16	0.10	0.35
QKF_1	3.89	7.86	3.81	0.20	0.21	0.33
QKF_2	3.88	7.81	3.76	0.18	0.15	0.33
05/01/2007	$\hat{\Sigma}_{11}$	$\hat{\Sigma}_{22}$	$\hat{\Sigma}_{33}$	$\hat{\rho}_{12}$	$\hat{\rho}_{13}$	$\hat{\rho}_{23}$
POL	2.73	9.79	3.70	0.29	0.44	0.34
QML_1	2.73	9.38	3.45	0.17	0.32	0.30
QML_2	2.72	9.43	3.41	0.14	0.27	0.32
QKF_1	3.22	14.22	5.50	0.13	0.27	0.28
QKF_2	3.21	14.20	5.47	0.13	0.26	0.28

Note: Values above for $\hat{\Sigma}_{ii}$ $(i=1,2,3)$ are 100 times the actual values, 04/01/2007 means January 4th, 2007, similar for 05/01/2007. Original data are synchronized by refresh time scheme for the POL and QML approaches. QKF is our new estimator obtained by combining the QML approach and Kalman filter together. QML is the estimator developed in Liu and Tang (2012) and POL is the estimator derived in Aït-Sahalia, Fan, and Xiu (2010). QML_1, and QKF_1 are obtained by assuming the covariance matrix of noises is a general covariance matrix. QML_2, and QKF_2 are obtained by assuming the covariance matrix of noises is a diagonal covariance matrix.

trading records at different time points of IBM, DELL and Microsoft are 10,043, 10,884, and 12,316 respectively on January 4th, and are 9,832, 10,935, and 11,939 respectively on January 5th. After synchronization, the sample size of synchronized data is 6,536 on January 4th and 6,142 on January 5th. The differences between methods may be due to the fact that the data synchronization has "smoothed" the price process, which may have bi-fold impacts—it may have reduced the effect from the jump of the prices or it simply has eliminated some informative dynamics. It is worthwhile to further investigate on the impact due to data synchronization.

In addition, we also see that the QML approach and the QKF approach obtain different estimates by treating the covariance matrix $\mathbf{A}$ of the microstructure noise differently, especially for the correlation estimates. For example, the correlation between IBM and MFT is estimated as 0.19 and 0.21 respectively by the QML and QKL approaches if $\mathbf{A}$ is considered as a general covariance matrix. While the estimates become 0.10 and 0.15 if $\mathbf{A}$ is considered as diagonal. Table 11 reports the estimated covariance and correlations of the microstructure noises. Though in general, correlations in Table 11 are small, some correlations take moderate values. As shown in our simulations, if the matrix $\mathbf{A}$ is misspecified, there can be a bias incurring in the integrated covariance matrix estimation.

4. CONCLUSION

We consider a state space model for high frequency financial trading data. The state space model can naturally han-

Table 11. Estimators for the elements of the covariance matrix of noises contaminated in observations of IBM, DELL and MFT $[\hat{A}_{ii}]$ times 10^7 and the correlations $\breve{\rho}^*_{ij}$ $(i,j=1,2,3)$ for empirical study in Section 3.2

04/01/2007	$\hat{A}_{11}$	$\hat{A}_{22}$	$\hat{A}_{33}$	$\breve{\rho}^*_{12}$	$\breve{\rho}^*_{13}$	$\breve{\rho}^*_{23}$
POL	0.06	3.50	1.28	0.08^*	0.17^*	0.01^*
QML_1	0.06	3.50	1.30	0.05^*	0.15^*	0.01^*
QKF_1	0.07	3.57	1.25	0.03^*	0.13^*	0.01
05/01/2007	$\hat{A}_{11}$	$\hat{A}_{22}$	$\hat{A}_{33}$	$\hat{A}_{12}$	$\hat{A}_{13}$	$\hat{A}_{23}$
POL	0.03	22.7	3.69	0.17^*	0.23^*	0.01
QML_1	0.03	22.8	3.72	0.04^*	0.10^*	0.02
QKF_1	0.04	22.0	4.18	0.01^*	0.03^*	0.00^*

Note: 04/01/2007 means January 4th, 2007, similar for 05/01/2007. Original data are synchronized by refresh time scheme for the POL and QML approaches. QKF_1 is our new estimator obtained by combining the QML approach and Kalman filter together. QML_1 is the estimator developed in Liu and Tang (2012) and POL_1 is the estimator derived in Aït-Sahalia, Fan, and Xiu (2010). Values with stars at right head are minus.

dle asynchronous observations by considering the problem in a scenario of missing data. We develop an EM algorithm for estimating the integrate covariance matrix with a general assumption on the covariance matrix of the microstructure noises. We show that the state space model approach performs promisingly in various scenarios.

A few problems remain open for further investigation. First, how to quantify and practically assess the impact due to the data asynchronicity is generally difficult because the mechanism lead to asynchronous data is complicated. The existing study generally assumes a special structure on the dependence between the price process and the observation times. Second, structural information from the market such as an industrial segment might be helpful for enhancing the performances of the estimating approaches. A study on how to incorporate such information in the framework of high frequency financial data analysis will be beneficial. Third, it is of great interest to investigate methods for assessing the level of uncertainties associated with the methods when dealing with asynchronous data.

APPENDIX

Lemma A.1. *Let* $\tilde{\mathbf{A}}_j = \left(\begin{smallmatrix} \mathbf{A}_{11j} & 0 \\ 0 & \mathbf{I}_{d-d_j} \end{smallmatrix}\right)$ *with* $\mathbf{A}_{11j}$ *to be the upper-left* $d_j \times d_j$ *block matrix of* $\mathbf{A}_j = \mathbf{B}_j \mathbf{A} \mathbf{B}_j' = \left(\begin{smallmatrix} \mathbf{A}_{11j} & \mathbf{A}_{12j} \\ \mathbf{A}_{21j} & \mathbf{A}_{22j} \end{smallmatrix}\right)$. *Giving the initial conditions* $\mathbf{X}_0 = \boldsymbol{\mu}_*$ *and* $\mathbf{P}_0^0 = \boldsymbol{\Sigma}_*$, *we have the following three results for the state space model (3):*

(1) *(Filtering). For* $j = 1, 2, \ldots, n$,

$$\mathbf{X}_j^{j-1} = \mathbf{X}_{j-1}^{j-1}, \tag{11}$$

$$\mathbf{P}_j^{j-1} = \mathbf{P}_{j-1}^{j-1} + \boldsymbol{\Sigma}\Delta_j, \tag{12}$$

$$\mathbf{X}_j^{j} = \mathbf{X}_j^{j-1} + \mathbf{K}_j(\tilde{\mathbf{Y}}_j - \tilde{\mathbf{B}}_j \mathbf{X}_j^{j-1}), \tag{13}$$

$$\mathbf{P}_j^{j} = (\mathbf{I} - \mathbf{K}_j \tilde{\mathbf{B}}_j)\mathbf{P}_j^{j-1}, \tag{14}$$

where $\tilde{\mathbf{Y}}_j = \left(\begin{smallmatrix} \mathbf{Y}_j^{(1)} \\ \mathbf{0} \end{smallmatrix}\right)$, $\tilde{\mathbf{B}}_j = \left(\begin{smallmatrix} \mathbf{B}_j^{(1)} \\ \mathbf{0} \end{smallmatrix}\right)$, *and*

$$\mathbf{K}_j = \mathbf{P}_j^{j-1} \tilde{\mathbf{B}}_j' (\tilde{\mathbf{B}}_j \mathbf{P}_j^{j-1} \tilde{\mathbf{B}}_j' + \tilde{\mathbf{A}}_j)^{-1}. \tag{15}$$

(2) *(Smoothing). For* $j = n, n-1, \ldots, 1$,

$$\mathbf{X}_{j-1}^{n} = \mathbf{X}_{j-1}^{j-1} + \mathbf{J}_{j-1}(\mathbf{X}_j^n - \mathbf{X}_j^{j-1}), \tag{16}$$

$$\mathbf{P}_{j-1}^{n} = \mathbf{P}_{j-1}^{j-1} + \mathbf{J}_{j-1}(\mathbf{P}_j^n - \mathbf{P}_j^{j-1})\mathbf{J}_{j-1}', \tag{17}$$

where

$$\mathbf{J}_{j-1} = \mathbf{P}_{j-1}^{j-1}(\mathbf{P}_j^{j-1})^{-1}. \tag{18}$$

(3)*. Under the initial condition*

$$\mathbf{P}_{n,n-1}^{n} = (\mathbf{I} - \mathbf{K}_n \tilde{\mathbf{B}}_n)\mathbf{P}_{n-1}^{n-1}, \tag{19}$$

the lag-one covariance smoother can also be derived as following: for $j = n, n-1, \ldots, 2$,

$$\mathbf{P}_{j-1,j-2}^{n} = \mathbf{P}_{j-1}^{j-1}\mathbf{J}_{j-2}' + \mathbf{J}_{j-1}(\mathbf{P}_{j,j-1}^{n} - \mathbf{P}_{j-1}^{j-1})\mathbf{J}_{j-2}'. \tag{20}$$

Lemma A.1 is an extension of Properties 6.1, 6.2 and 6.3 in Shumway and Stoffer (2006).

Proof of Lemma A.1. Since Kalman filter is based on updating the conditional expectation under increased information and previous information. Therefore, we firstly give the following well-know conclusion.

If

$$\begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}\right),$$

then

$$\begin{aligned} &\mathbf{W}_2 | \mathbf{W}_1 = w_1 \\ &\sim N\left(\boldsymbol{\mu}_2 + \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}(w_1 - \boldsymbol{\mu}_1), \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\right). \end{aligned} \tag{21}$$

Denote that

$$\gamma_j^s = \{\mathbf{V}_j, \mathbf{V}_{j+1}, \ldots, \mathbf{V}_s, \mathbf{U}_{j+1}, \mathbf{U}_{j+2}, \ldots, \mathbf{U}_s\},$$

where $s > j$. Then for $h < j$, we can find that y_h, $\mathbf{V}_h$, and $\boldsymbol{\epsilon}_h = \tilde{\mathbf{Y}}_h - E(\tilde{\mathbf{Y}}_h | y_{h-1}) = \tilde{\mathbf{Y}}_h - \tilde{\mathbf{B}}_h \mathbf{X}_h^{h-1}$ are independent of γ_j^s, and $\tilde{\mathbf{Y}}_h$ and $\boldsymbol{\epsilon}_j$ are independent as $\boldsymbol{\epsilon}_j$ and $\tilde{\mathbf{Y}}_j$ are Gaussian random variables with covariance

$$\begin{aligned} E(\boldsymbol{\epsilon}_j \tilde{\mathbf{Y}}_h') &= E\left\{\tilde{\mathbf{Y}}_j \tilde{\mathbf{Y}}_h' - E(\tilde{\mathbf{Y}}_j | y_{j-1})\tilde{\mathbf{Y}}_h\right\} \\ &= E\left\{\tilde{\mathbf{Y}}_j \tilde{\mathbf{Y}}_h' - E(\tilde{\mathbf{Y}}_j \tilde{\mathbf{Y}}_h' | y_{j-1})\right\} = 0. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbf{X}_j^{j-1} &= E(\mathbf{X}_j | y_{j-1}) = E(\mathbf{X}_{j-1} + \mathbf{V}_j | y_{j-1}) \\ &= E(\mathbf{X}_{j-1} | y_{j-1}) = \mathbf{X}_{j-1}^{j-1}, \\ \mathbf{P}_j^{j-1} &= E\left\{(\mathbf{X}_j - \mathbf{X}_j^{j-1})(\mathbf{X}_j - \mathbf{X}_j^{j-1})'\right\} \end{aligned}$$

$$
\begin{aligned}
&= E\left\{(\mathbf{X}_{j-1} - \mathbf{X}_j^{j-1})(\mathbf{X}_{j-1} - \mathbf{X}_j^{j-1})'\right\} \\
&\quad + E(\mathbf{V}_j\mathbf{V}_j') + E\left\{(\mathbf{X}_{j-1} - \mathbf{X}_j^{j-1})\mathbf{V}_j'\right\} \\
&\quad + E\left\{\mathbf{V}_j(\mathbf{X}_{j-1} - \mathbf{X}_j^{j-1})'\right\} \\
&= \mathbf{P}_{j-1}^{j-1} + \mathbf{\Sigma}\Delta_j,
\end{aligned}
$$

which are (11) and (12). Next, we prove (13). To prove it, we firstly derive the joint distribution of $\mathbf{X}_j$ and $\boldsymbol{\epsilon}_j$ conditional on y_{j-1}. We have

$$
\begin{aligned}
&\mathrm{Var}(\boldsymbol{\epsilon}_j|y_{j-1}) \\
&= \mathrm{Var}(\boldsymbol{\epsilon}_j) = \mathrm{Var}\{\tilde{\mathbf{B}}_j(\mathbf{X}_j - \mathbf{X}_j^{j-1}) + \tilde{\mathbf{U}}_j\} \\
&= \tilde{\mathbf{B}}_j\mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j' + \tilde{\mathbf{A}}_j, \\
&\mathrm{Cov}(\mathbf{X}_j, \boldsymbol{\epsilon}_j|y_{j-1}) \\
&= \mathrm{Cov}(\mathbf{X}_j, \tilde{\mathbf{Y}}_j - \tilde{\mathbf{B}}_j\mathbf{X}_j^{j-1}|y_{j-1}) \\
&= E\left(\left(\mathbf{X}_j - \mathbf{X}_j^{j-1}\right)\left\{\tilde{\mathbf{B}}_j(\mathbf{X}_j - \mathbf{X}_j^{j-1}) + \mathbf{U}_j\right\}'|y_{j-1}\right) \\
&= \mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j'.
\end{aligned}
$$

The last equation is because $\mathbf{X}_j$, $\mathbf{X}_j^{j-1}$ are independent of $\tilde{\mathbf{U}}_j$ and $E\{\mathbf{X}_j^{j-1}(\mathbf{X}_j - \mathbf{X}_j^{j-1})|y_{j-1}\} = 0$. Then the joint distribution of $\mathbf{X}_j$ and $\boldsymbol{\epsilon}_j$ conditional on y_{j-1} follows

$$
\begin{aligned}
&\begin{pmatrix} \mathbf{X}_j \\ \boldsymbol{\epsilon}_j \end{pmatrix}|y_{j-1} \\
&\sim N\left(\begin{pmatrix} \mathbf{X}_j^{j-1} \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{P}_j^{j-1} & \mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j' \\ \tilde{\mathbf{B}}_j'\mathbf{P}_j^{j-1} & \tilde{\mathbf{B}}_j\mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j' + \tilde{\mathbf{A}}_j \end{pmatrix}\right).
\end{aligned}
$$

Therefore, by (21) we have that

$$
\begin{aligned}
(22) \qquad \mathbf{X}_j^j &= E(\mathbf{X}_j|y_{j-1}, \mathbf{Y}_j^{(1)}) = E(\mathbf{X}_j|y_{j-1}, \boldsymbol{\epsilon}_j) \\
&= \mathbf{X}_j^{j-1} + \mathbf{K}_j\boldsymbol{\epsilon}_j \\
&= \mathbf{X}_j^{j-1} + \mathbf{K}_j(\tilde{\mathbf{Y}}_j - \tilde{\mathbf{B}}_j\mathbf{X}_j^{j-1}) \\
&= \mathbf{X}_j^{j-1} + \mathbf{K}_j\tilde{\mathbf{B}}_j(\mathbf{X}_j - \mathbf{X}_j^{j-1}) + \mathbf{K}_j\tilde{\mathbf{U}}_j,
\end{aligned}
$$

where

$$
\mathbf{K}_j = \mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j'(\tilde{\mathbf{B}}_j\mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j' + \tilde{\mathbf{A}}_j)^{-1}.
$$

And,

$$
\begin{aligned}
\mathbf{P}_j^j &= \mathrm{Var}(\mathbf{X}_j|y_{j-1}, \boldsymbol{\epsilon}_j) \\
&= \mathbf{P}_j^{j-1} - \mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j'(\tilde{\mathbf{B}}_j\mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j' + \tilde{\mathbf{A}}_j)^{-1}\tilde{\mathbf{B}}_j\mathbf{P}_j^{j-1} \\
&= (\mathbf{I} - \mathbf{K}_j\tilde{\mathbf{B}}_j)\mathbf{P}_j^{j-1},
\end{aligned}
$$

which is (15). Therefore, we have proven the conclusion for Kalman filter part of Lemma A.1.

Next, we prove the smoothing part of Lemma A.1. Since y_{j-1}, $\mathbf{X}_j - \mathbf{X}_j^{j-1}$ and γ_j^s are mutually independent. Therefore, by applying (21), we have that

$$
\begin{aligned}
&E(\mathbf{X}_{j-1}|y_{j-1}, \mathbf{X}_j - \mathbf{X}_j^{j-1}, \gamma_j^s) \\
&= E(\mathbf{X}_{j-1}|y_{j-1}, \mathbf{X}_j - \mathbf{X}_j^{j-1}) \\
&= \mathbf{X}_{j-1}^{j-1} + \mathbf{J}_{j-1}(\mathbf{X}_j - \mathbf{X}_j^{j-1}),
\end{aligned}
$$

where the first equation is because $\mathbf{X}_{j-1}$ is independent of γ_j^s. By applying (11), we have

$$
\begin{aligned}
&\mathbf{J}_{j-1} \\
&= \mathrm{Cov}(\mathbf{X}_{j-1}, \mathbf{X}_j - \mathbf{X}_j^{j-1}|y_{j-1})\left\{\mathrm{Var}\left(\mathbf{X}_j - \mathbf{X}_j^{j-1}|y_{j-1}\right)\right\}^{-1} \\
&= \mathbf{P}_{j-1}^{j-1}(\mathbf{P}_j^{j-1})^{-1}
\end{aligned}
$$

where the last equation is because

$$
\begin{aligned}
(23) \quad &\mathbf{P}_{j,j-1}^{j-1} \\
&= E\{(\mathbf{X}_j - \mathbf{X}_j^{j-1})(\mathbf{X}_{j-1} - \mathbf{X}_{j-1}^{j-1})'\} \\
&= E\{(\mathbf{X}_{j-1} + \mathbf{V}_j - \mathbf{X}_{j-1}^{j-1})(\mathbf{X}_{j-1} - \mathbf{X}_{j-1}^{j-1})'\} = \mathbf{P}_{j-1}^{j-1}.
\end{aligned}
$$

Then (16) are proven by

$$
\begin{aligned}
\mathbf{X}_{j-1}^n &= E\left(\mathbf{X}_{j-1}|y_n\right) \\
&= E\left(E\left(\mathbf{X}_{j-1}|y_{j-1}, \mathbf{X}_j - \mathbf{X}_j^{j-1}, \gamma_j^n\right)|y_n\right) \\
&= \mathbf{X}_{j-1}^{j-1} + \mathbf{J}_{j-1}(\mathbf{X}_j^n - \mathbf{X}_j^{j-1}).
\end{aligned}
$$

Therefore

$$
\begin{aligned}
(24) \quad &\mathbf{X}_{j-1} - \mathbf{X}_{j-1}^n + \mathbf{J}_{j-1}\mathbf{X}_j^n = \mathbf{X}_{j-1} - \mathbf{X}_{j-1}^{j-1} + \mathbf{J}_{j-1}\mathbf{X}_{j-1}^{j-1}.
\end{aligned}
$$

Hence by taking expectation of each side of (24) multiplied by the transpose of itself, we have

$$
\begin{aligned}
(25) \qquad &\mathbf{P}_{j-1}^n + \mathbf{J}_{j-1}E(\mathbf{X}_j^n\mathbf{X}_j^{n'})\mathbf{J}_{j-1}' \\
&= \mathbf{P}_{j-1}^{j-1} + \mathbf{J}_{j-1}E(\mathbf{X}_{j-1}^{j-1}\mathbf{X}_{j-1}^{j-1'})\mathbf{J}_{j-1}'
\end{aligned}
$$

where the last equation is because $E\{\mathbf{X}_j^n(\mathbf{X}_{j-1} - \mathbf{X}_{j-1}^n)'\} = 0$ and $E\{\mathbf{X}_{j-1}^{j-1}(\mathbf{X}_{j-1} - \mathbf{X}_{j-1}^{j-1})'\} = 0$, which can be obtained by firstly denoting

$$
\tilde{\mathbf{X}}_j^h = \mathbf{X}_h - \mathbf{X}_j^h,
$$

and then for $h \le l$, $h \le i$ and $l \le j$,

$$
\begin{aligned}
(26) \qquad &E\left(\mathbf{V}_j\tilde{\mathbf{X}}_j^{l'}\right) = 0, \\
&E\left(\mathbf{X}_i^h\tilde{\mathbf{X}}_j^{l'}\right) = E\left\{\mathbf{X}_i^h(\mathbf{X}_j - \mathbf{X}_j^l)'\right\} \\
&\qquad = E\left\{E\left(\mathbf{X}_i^h\mathbf{X}_j'|y_l\right)\right\} - E\left(\mathbf{X}_i^h\mathbf{X}_j^{l'}\right) \\
&\qquad = E\left(\mathbf{X}_i^h\mathbf{X}_j^{l'}\right) - E\left(\mathbf{X}_i^h\mathbf{X}_j^{l'}\right) = 0,
\end{aligned}
$$

and for $h \ge l$, $h \le i$ and $l \le j$,

$$E\left(\mathbf{X}_i^h\tilde{\mathbf{X}}_j^{l'}\right) = E\left\{\mathbf{X}_i^h(\mathbf{X}_j - \mathbf{X}_j^l)'\right\}$$
$$= E\left\{E\left(\mathbf{X}_i^h\mathbf{X}_j'|y_h\right)\right\} - E\left(\mathbf{X}_i^h\mathbf{X}_j^{l'}\right)$$
$$= E\left(\mathbf{X}_i^h\mathbf{X}_j^{h'}\right) - E\left(\mathbf{X}_i^h\mathbf{X}_j^{l'}\right) \neq 0.$$

On the other hand, since

$$E(\mathbf{X}_j^n\mathbf{X}_j^{n'}) = E(\mathbf{X}_j\mathbf{X}_j') - \mathbf{P}_j^n = E(\mathbf{X}_{j-1}\mathbf{X}_{j-1}') + \mathbf{\Sigma}\Delta_j - \mathbf{P}_j^n$$
$$= E(\mathbf{X}_{j-1}^{j-1}\mathbf{X}_{j-1}^{j-1'}) + \mathbf{P}_{j-1}^{j-1} + \mathbf{\Sigma}\Delta_j - \mathbf{P}_j^n$$
$$= E(\mathbf{X}_{j-1}^{j-1}\mathbf{X}_{j-1}^{j-1'}) + \mathbf{P}_j^{j-1} - \mathbf{P}_j^n$$

by realizing $\mathbf{V}_j$ is independent of $\mathbf{X}_{j-1}$ and (12), therefore we obtain equation (17) by combining (25) and above equation.

The lag-one covariance smoother can also be proven by direct calculation. By (11) and (22) we have

$$\mathbf{P}_{j,j-1}^j = E\left(\tilde{\mathbf{X}}_j^j\tilde{\mathbf{X}}_{j-1}^j\right)$$
$$= E\left(\left\{\tilde{\mathbf{X}}_j^{j-1} - \mathbf{K}_j(\tilde{\mathbf{B}}_j\tilde{\mathbf{X}}_j^{j-1} + \tilde{\mathbf{U}}_j)\right\}\right.$$
$$\left.\left\{\tilde{\mathbf{X}}_{j-1}^{j-1} - \mathbf{J}_{j-1}\mathbf{K}_j(\tilde{\mathbf{B}}_j\tilde{\mathbf{X}}_j^{j-1} + \tilde{\mathbf{U}}_j)\right\}'\right)$$
$$= \mathbf{P}_{j,j-1}^{j-1} - \mathbf{K}_j\tilde{\mathbf{B}}_j\mathbf{P}_{j,j-1}^{j-1} - \mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j'\mathbf{K}_j'\mathbf{J}_{j-1}'$$
$$+ \mathbf{K}_j(\tilde{\mathbf{B}}_j\mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j' + \tilde{\mathbf{A}}_j)\mathbf{K}_j'\mathbf{J}_{j-1}'$$
$$= \mathbf{P}_{j-1}^{j-1} - \mathbf{K}_j\tilde{\mathbf{B}}_j\mathbf{P}_{j-1}^{j-1} - \mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j'\mathbf{K}_j'\mathbf{J}_{j-1}'$$
$$+ \mathbf{P}_j^{j-1}\tilde{\mathbf{B}}_j'\mathbf{K}_j'\mathbf{J}_{j-1}'$$
$$= (\mathbf{I} - \mathbf{K}_j\tilde{\mathbf{B}}_j)\mathbf{P}_{j-1}^{j-1}.$$

The fourth equation is because (23) and (15). Therefore (19) is proven by above and letting $j = n$.

To prove (20), we reuse (24) to have

$$(\tilde{\mathbf{X}}_{j-1}^n + \mathbf{J}_{j-1}\mathbf{X}_j^n)(\tilde{\mathbf{X}}_{j-2}^n + \mathbf{J}_{j-2}\mathbf{X}_{j-1}^n)' \tag{27}$$
$$= (\tilde{\mathbf{X}}_{j-1}^{j-1} + \mathbf{J}_{j-1}\mathbf{X}_{j-1}^{j-1})(\tilde{\mathbf{X}}_{j-2}^{j-2} + \mathbf{J}_{j-2}\mathbf{X}_{j-2}^{j-2})'.$$

And, on the other hand, we have

$$E\left(\mathbf{X}_j^n\tilde{\mathbf{X}}_{j-2}^{n'}\right) = 0, E\left(\tilde{\mathbf{X}}_{j-1}^n\mathbf{X}_{j-1}^{n'}\right) = 0, E\left(\tilde{\mathbf{X}}_{j-1}^{j-1}\mathbf{X}_{j-2}^{j-2'}\right) = 0$$

by (26). Combining above three equations, (16), (27), and $\tilde{\mathbf{X}}_{j-1}^{j-1} = (\mathbf{I} - \mathbf{K}_{j-1}\tilde{\mathbf{B}}_{j-1})\tilde{\mathbf{X}}_{j-1}^{j-2} + \mathbf{K}_{j-1}\tilde{\mathbf{U}}_{j-1}$ obtained from (22), we have

$$\mathbf{P}_{j-1,j-2}^n$$
$$= E\left\{\left(\tilde{\mathbf{X}}_{j-1}^{j-1} + \mathbf{J}_{j-1}\mathbf{X}_{j-1}^{j-1}\right)\left(\tilde{\mathbf{X}}_{j-2}^{j-2} + \mathbf{J}_{j-2}\mathbf{X}_{j-2}^{j-2}\right)'\right\}$$
$$- \mathbf{J}_{j-1}E(\mathbf{X}_j^n\mathbf{X}_{j-1}^{n'})\mathbf{J}_{j-2}'$$
$$= (\mathbf{I} - \mathbf{K}_{j-1}\tilde{\mathbf{B}}_{j-1})\mathbf{P}_{j-1,j-2}^{j-2} + \mathbf{J}_{j-1}\mathbf{K}_{j-1}\tilde{\mathbf{B}}_{j-1}\mathbf{P}_{j-1,j-2}^{j-2}$$
$$+ \mathbf{J}_{j-1}\left\{E(\mathbf{X}_{j-1}^{j-1}\mathbf{X}_{j-2}^{j-2'}) - E(\mathbf{X}_j^n\mathbf{X}_{j-1}^{n'})\right\}\mathbf{J}_{j-2}'$$
$$= \mathbf{P}_{j-1}^{j-1}\mathbf{J}_{j-2}' + \mathbf{J}_{j-1}\mathbf{K}_{j-1}\tilde{\mathbf{B}}_{j-1}\mathbf{P}_{j-1,j-2}^{j-2}$$
$$+ \mathbf{J}_{j-1}\left\{E(\mathbf{X}_{j-1}^{j-1}\mathbf{X}_{j-2}^{j-2'}) - E(\mathbf{X}_j^n\mathbf{X}_{j-1}^{n'})\right\}\mathbf{J}_{j-2}'$$

as $\mathbf{P}_{j-1}^{j-1} = (I - \mathbf{K}_{j-1}\tilde{\mathbf{B}}_{j-1})\mathbf{P}_{j-2}^{j-1}$, $\mathbf{J}_{j-2} = \mathbf{P}_{j-2}^{j-2}(\mathbf{P}_{j-1}^{j-2})^{-1}$ and (23). Moreover, since

$$E(\mathbf{X}_{j-1}^{j-1}\mathbf{X}_{j-2}^{j-2'}) - E(\mathbf{X}_j^n\mathbf{X}_{j-1}^{n'})$$
$$= E(\mathbf{X}_{j-1}^{j-2}\mathbf{X}_{j-2}^{j-2'}) - \{E(\mathbf{X}_j\mathbf{X}_{j-1}') - \mathbf{P}_{j,j-1}^n\}$$
$$= \left\{E(\mathbf{X}_{j-1}\mathbf{X}_{j-2}') - \mathbf{P}_{j-1,j-2}^{j-2}\right\}$$
$$- \left\{E(\mathbf{X}_{j-1}\mathbf{X}_{j-2}') + \mathbf{\Sigma}\Delta_{j-1} - \mathbf{P}_{j,j-1}^n\right\}$$
$$= \mathbf{P}_{j,j-1}^n - \left(\mathbf{P}_{j-2}^{j-2} + \mathbf{\Sigma}\Delta_{j-1}\right)$$

by (13) and (23), therefore by (12) and (14) we have

$$\mathbf{P}_{j-1,j-2}^n = \mathbf{P}_{j-1}^{j-1}\mathbf{J}_{j-2}' + \mathbf{J}_{j-1}\left\{\mathbf{P}_{j,j-1}^n -\right.$$
$$\left.\left(\mathbf{P}_{j-1}^{j-2} - \mathbf{K}_{j-1}\tilde{\mathbf{B}}_{j-1}\mathbf{P}_{j-2}^{j-1}(\mathbf{J}_{j-2}')^{-1}\right)\right\}\mathbf{J}_{j-2}'$$
$$= \mathbf{P}_{j-1}^{j-1}\mathbf{J}_{j-2}' + \mathbf{J}_{j-1}\left(\mathbf{P}_{j,j-1}^n - \mathbf{P}_{j-1}^{j-1}\right)\mathbf{J}_{j-2}',$$

which is actually (20). Therefore, we finished the proof of Lemma A.1. □

Received 15 June 2013

REFERENCES

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies* **18**, 315–416.

Aït-Sahalia, Y., and Mykland, P. A. (2009). Estimating volatility in the presence of market microstructure noise: A review of the theory and practical considerations. *Handbook of Financial Time Series*, Thomas Mikosch et al., eds, Springer-Verlag.

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2011). Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics* **160**, 160–175. MR2745875

Aït-Sahalia, Y., Fan, J., and Xiu, D. (2010). High frequency covariance estimates with noisy and nonsynchronous financial data. *Journal of the American Statistical Association* **105**, 1505–1517.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica* **71**, 579–625. MR1958138

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2008). Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility. *Review of Economics and Statistics* **89**, 701–720.

Barndorff-Nielsen, O. E., and Shephard, N. (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica* **72**, 885–925. MR2051439

Barndorff-Nielsen, O. E., and Shephard, N. (2007). Variation, jumps and high frequency data in financial econometrics. In: *Advances in Economics and Econometrics. Theory and Applications*, R. Blundell, T. Persson, and W. K. Newey, eds, Ninth World Congress, Econometric Society Monographs, Cambridge University Press, 328–372.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* **76**, 1481–1536. MR2468558

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2011). Multivariate realized kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics* **162**, 149–169. MR2795610

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* **81**, 637–654.

Christensen, K., Kinnebrock, S., and Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics* **159**, 116–133. MR2720847

Corsi, F., Peluso, S. and Audrino, F. (2012). Missing in asynchronicity: A Kalman-EM approach for multivariate realized covariance estimation (manuscript).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society, Series B* **39**, 1–38. MR0501537

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica* **50**, 987–1007. MR0666121

Engle, R. F., and Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* **66**, 1127–1162. MR1639411

Fan, J., and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association* **102**, 1349–1362. MR2372538

Gloter, A., and Jacod, J. (2001). Diffusions with measurement errors: I. Local asymptotic normality. *European Series in Applied and Industrial Mathematics* **5**, 225–242. MR1875672

Hansen, P. R., and Lunde, A. (2006). Realized variance and market microstructure noise (with discussion). *Journal of Business & Economic Statistics* **24**, 127–218. MR2234447

Harris, F., McInish, T., Shoesmith, G. and Wood, R. (1995). Cointegration, error correction and price discovery on informationally linked security markets. *Journal of Financial and Quantitative Analysis* **30**, 563–581.

Hoshikawa, T., Kanatani, T., Nagai, K. and Nashiyama, Y. (2008). Nonparametric estimation methods of integrated multivariate volatilities. *Econometric Review* **27**, 112–138. MR2424809

Jacod, J., Li, Y., Mykland, P., Podolskij, M., and Vetter, M. (2009). Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Processes and Their Applications* **119**, 2249–2276. MR2531091

Li, Y., Mykland, P. A., Renault, E., Zhang, L., and Zheng, X. (2009). Realized volatility when endogeneity of time matters (manuscript).

Liu, C., and Tang, C. Y. (2012). A quasi-maximum likelihood approach for integrated covariance matrix estimation with high frequency data (manuscript).

Malliavin, P., and Mancino, M. (2002). Fourier series method for measurement of multivariate volatilities. *Finance and Stochastics* **6**, 49–61. MR1885583

Malliavin, P., and Mancino, M.(2009). A Fourier transform method for nonparametric estimation of multivariate volatility. *The Annals of Statistics* **37**, 1983–2010. MR2533477

Mykland, P. A., and Zhang, L. (2010). The econometrics of high frequency data. *Statistical Methods for Stochastic Differential Equations*, M. Kessler, A. Lindner, and M. Sorensen, eds, Chapman & Hall/CRC Press, forthcoming. MR2976983

Peluso, S., Corsi, F. and Mira, A. (2012). A bayesian high-frequency estimator of the multivariate covariance of noisy and asynchronous returns. Available at SSRN: http://ssrn.com/abstract=2003492 or http://dx.doi.org/10.2139/ssrn.2003492.

Shephard N., and Xiu, D. (2012). Econometric analysis of multivariate realised QML: Estimation of the covariation of equity prices under asynchronous trading (manuscript).

Shumway, R. and Stoffer, D. S. (2011). *Time Series Analysis and Its Application: With R Examples* (3rd ed.). New York: Springer. MR2228626

Tao, M., Wang, Y., Yao, Q., and Zou, J. (2011). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *Journal of the American Statistical Association* **106**, 1025–1040. MR2894761

Wang, Y., and Zou, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *The Annals of Statistics* **38**, 943–978. MR2604708

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11**, 95–103. MR0684867

Xiu, D. (2010). Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics* **159**, 235–250. MR2720855

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high frequency data. *Journal of the American Statistical Association* **100**, 1394–1411. MR2236450

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli* **12**, 1019–1043. MR2274854

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics* **160**, 33–47. MR2745865

Cheng Liu
Sim Kee Boon Institute for Financial Economics
Singapore Management University
90 Stamford Road, 178903
Singapore
E-mail address: chengliu@smu.edu.sg

Cheng Yong Tang
Business School
University of Colorado Denver
Campus Box 165, PO Box 173364
Denver CO 80217-3364
USA
E-mail address: chengyong.tang@ucdenver.edu