Statistica Sinica Preprint No: SS-2017-0489	
Title	Efficient and positive semidefinite pre-averaging realized covariance estimator
Manuscript ID	SS-2017-0489
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0489
Complete List of Authors	Liang-Ching Lin Ying Chen Guangming Pan and Vladimir Spokoiny
Corresponding Author	Liang-Ching Lin
E-mail	lclin@mail.ncku.edu.tw
Notice: Accepted version subject to English editing.	

Statistica Sinica

Efficient and positive semidefinite pre-averaging realized covariance estimator

Liang-Ching Lin[1], Ying Chen[2], Guangming Pan[3] and Vladimir Spokoiny[4]

[1] *National Cheng Kung University,* [2]*National University of Singapore,*
[3]*Nanyang Technological University and* [4]*Weierstrass Institute*

Abstract: We propose a realized-covariance estimator based on efficient multiple pre-averaging (EMP) for asynchronous and noisy high-frequency data. The EMP estimator is consistent, guaranteed to be positive-semidefinite, and achieves the optimal convergence rate at $n^{-1/4}$. It is constructed based on 1) an innovative synchronizing technique utilizing all of the available price information and 2) an eigenvalue correction method to ensure positive-semidefiniteness without sacrificing the optimal convergence rate. A simulation study demonstrates the good performance of the EMP estimator for finite samples in terms of accuracy, properties, and convergence rate. In real data analysis, the EMP covariance estimator delivers more stable performance compared with alternative estimators. The new estimator also outperforms alternative realized-covariance estimators in terms of portfolio selection.

Key words and phrases: Asynchronous and noisy high-frequency data, eigenvalue correction, synchronizing technique.

1. Introduction

Covariance plays an important role in portfolio allocation, derivative pricing, hedging, risk management, and many other modern financial applications. Estimating covariance has been of great interest to academics and industry practitioners alike. Realized covariance, a model-free estimator, has attracted attention given the availability of large-scale intra-daily data sampled at second, millisecond, or even nanosecond frequency. Realized covariance is quantified as a quadratic variation of the high-frequency data. It is theoretically consistent (Andersen and Bollerslev, 1998; Barndorff-Nielsen and Shephard, 2002a,b; Andersen et al., 2003) and also demonstrates good accuracy in numerous applications (French et al., 1987; Andersen and Bollerslev, 1998; Andersen et al., 2001).

A direct calculation of realized covariance from the high-frequency raw data is, however, inconsistent due to the existence of asynchronous trading and microstructure noise. Various synchronizing techniques are used to preprocess the asynchronous raw data. The most common approaches are the previous tick technique (Wasserfallen and Zimmermann, 1985; Dacorogna et al., 2001) and the refresh time technique (Barndorff-Nielsen et al., 2008; Hautsch et al., 2012; Aït-Sahalia et al., 2010). While the former may distort the dependence of the multiple price processes in the raw data, the latter may lead to low sample sizes if one or more assets are illiquid. Christensen et al. (2013) used the approach of Hayashi and Yoshida, which depends on the selection of a smoothing parameter. Corsi et al. (2015) and Shephard and Xiu (2017) proposed the Kalman filter technique, which is under the

Gaussian distributional assumption. It is worth noting that none of the existing techniques considers the intrinsic data features, e.g. negative serial correlation.

Microstructure noise hinders the synchronizing process. Bias correction thus becomes necessary, yet often at the cost of efficiency. Some important contributions to this problem include but are not limited to the multivariate scaled estimator (Zhang, 2011; Wang and Zou, 2010; Zhang et al., 2005), the multivariate realized kernel estimator (Barndorff-Nielsen et al., 2011, 2008; Zhou, 1996; Hansen and Lunde, 2006), and the quasi-maximum likelihood realized-covariance estimator (QMLE, Aït-Sahalia et al., 2010; Xiu, 2010). The multiple pre-averaging (MPA) estimator removes noise using a pre-averaging procedure, see Christensen et al. (2010), Jacod et al. (2009). While the QMLE and MPA estimators are $n^{-1/4}$-consistent, where $n^{-1/4}$ is the optimal convergence rate, the other two are suboptimal at $\mathcal{O}_p\left(n^{-1/6}\right)$ and $\mathcal{O}_p\left(n^{-1/5}\right)$, respectively. In addition, Shephard and Xiu (2017) developed a multivariate realized quasi-maximum likelihood estimator based on synchronized observations, which was positive-definite, $n^{-1/4}$-consistent, and asymptotically mixed-normal.

The bias correction approaches can introduce negative covariance estimators. The negative eigenvalues, though at small magnitude, may change the stochastic behavior of the covariance estimator and in some cases even its consistency. The MPA estimator, for instance, is not guaranteed to be positive-semidefinite. To enforce the right properties, the convergence rate becomes suboptimal. Several eigenvalue correction approached have been proposed. One can replace negative eigenvalues with small positive values (McNeil et al., 2005; Schaeffer, 2014) or zeros (Rebonato and Jäckel, 1999). Varneskov (2015) proposed an eigenvalue truncation procedure and showed that the correction was asymptotically negligible. Ikeda (2016) presented a Cholesky-type correction without altering the asymptotic distribution of the two-scale realized-kernel estimator.

In our study, we propose a realized-covariance estimator based on efficient multiple pre-averaging (EMP), which is consistent, positive-semidefinite, and simultaneously achieves the optimal $n^{-1/4}$ convergence rate. The EMP estimator benefits from two innovative approaches. We develop a synchronizing technique called high-frequency filtering (HFF) to recover the "missing" records of high-frequency data by learning from the dependence of the same price processes that are synchronously sampled at low frequency. Given a prior (realized) covariance estimator and the negative autocorrelation, the "unobserved" records in the asynchronous data are iteratively filtered. We also present an eigenvalue correction method for a consistent yet negative realized covariance estimator, where the negative eigenvalues with small magnitude are replaced with the absolute values to enforce positive-semidefiniteness of the estimator. We describe the convergence properties of the filtered high-frequency synchronous series. We show that the corrected realized-covariance estimator has the same limiting distribution as the consistent but negative-semidefinite realized-covariance estimator with the optimal convergence rate. Both approaches are model-free in that they require neither distributional assumptions nor tuning parameters. The approaches are general and can be used for any type of covariance and correlation estimator.

The remainder of this paper is organized as follows. Section 2 is the model setting. Section 3 presents the EMP realized-covariance estimator. We detail the HFF technique and the eigenvalue correction method and provide the asymptotic results. We demonstrate the finite sample performance of the EMP estimator via an extensive simulation study in Section 4. Empirical data analysis is conducted in Section 5. Section 6 provides concluding remarks. All of the theoretical proofs are contained in the supplementary material.

2. Model Setup

Consider p assets traded over a time interval $t \in [0, 1]$. The efficient log prices $\mathbf{X}_t \in \mathbb{R}^p$ are assumed to follow the Brownian semimartingale model,

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t^\top d\mathbf{B}_t, \quad t \in [0, 1], \tag{2.1}$$

where $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{pt})^\top$ is the drift vector of the multiple assets, $\mathbf{B}_t = (B_{1t}, \ldots, B_{pt})^\top$ is a standard p-dimensional Brownian motion, $\boldsymbol{\sigma}_t$ is a $p \times p$ matrix, and the symbol $\top$ represents the Hermitian transpose. The quadratic variation of $\boldsymbol{X}_t$ is given by:

$$[\boldsymbol{X}, \boldsymbol{X}]_t = \int_0^t \Sigma_u du = \int_0^t \boldsymbol{\sigma}_u^\top \boldsymbol{\sigma}_u du. \tag{2.2}$$

The integrated volatility matrix, denoted by Σ, is defined as:

$$\Sigma \equiv \int_0^1 \Sigma_u du = \int_0^1 \boldsymbol{\sigma}_u^\top \boldsymbol{\sigma}_u du. \tag{2.3}$$

Our goal is to estimate the integrated volatility matrix Σ given asynchronous and noisy data traded at high frequency.

The synchronous log prices $Y_{t_j} \in \mathbb{R}^p$ at discrete and regular time points $t_j = j/n$, $j = 0, \ldots, n$ are assumed to follow the continuous diffusion model with additive noise:

$$\mathbf{Y}_{t_j} = \mathbf{X}_{t_j} + \boldsymbol{\epsilon}_{t_j} \tag{2.4}$$

where $\mathbf{X}_{t_j} = (X_{1,t_j}, \ldots, X_{p,t_j})^\top$ denotes the efficient noise-free log prices, and $\boldsymbol{\epsilon}_{t_j}$ is an i.i.d. microstructure noise with zero mean and finite variance $E(\boldsymbol{\epsilon}_{\boldsymbol{t_j}} \boldsymbol{\epsilon}_{\boldsymbol{t_j}}^\top) = \text{diag}\{\eta_1^2, \ldots, \eta_p^2\}$. $\mathbf{X}_{t_j}$ are assumed to be mutually independent with $\boldsymbol{\epsilon}_{t_j}$.

Given the return series $R_{i,t_j} = Y_{i,t_j} - Y_{i,t_{j-1}}$, it is easy to show that there is negative lag-1 autocorrelation

$$\begin{aligned} \frac{Cov(R_{i,t_{j-1}}, R_{i,t_j})}{\sqrt{Var(R_{i,t_{j-1}}) Var(R_{i,t_j})}} &= \frac{-\eta_i^2}{\sqrt{\left(\frac{1}{n} E \int_{t_{j-2}}^{t_{j-1}} \Sigma_{ii,u} du + 2\eta_i^2\right)\left(\frac{1}{n} E \int_{t_{j-1}}^{t_j} \Sigma_{ii,u} du + 2\eta_i^2\right)}} \\ &\approx -0.5, \quad i = 1, \ldots, p, \quad j = 2, \ldots, n \end{aligned} \tag{2.5}$$

where $\Sigma_{ii,u}$ denotes the (i,i)-component of Σ_u in (2.2).

Given synchronous data, the MPA estimator (Christensen et al., 2010), denoted as S_1, is computed:

$$S_1 = \frac{n}{n-k_n+2}\frac{12}{k_n}\sum_{j=0}^{n-k_n+1}\bar{\mathbf{Y}}_{t_j}^n(\bar{\mathbf{Y}}_{t_j}^n)^\top - \frac{12}{2n\theta^2}\sum_{j=1}^{n}(\boldsymbol{Y}_{t_j}-\boldsymbol{Y}_{t_{j-1}})(\boldsymbol{Y}_{t_j}-\boldsymbol{Y}_{t_{j-1}})^\top, \quad (2.6)$$

where $\bar{\mathbf{Y}}_{t_j}^n = \frac{1}{k_n}\left(\sum_{\ell=k_n/2}^{k_n-1}\boldsymbol{Y}_{t_{j+\ell}} - \sum_{\ell=0}^{k_n/2}\boldsymbol{Y}_{t_{j+\ell}}\right)$, $k_n = \lfloor\theta\sqrt{n}\rfloor$ with a given constant $\theta > 0$ and the last term of (2.6) is a bias correction term. The MPA estimator is an unbiased estimator of Σ with convergence rate $\mathcal{O}_p\left(n^{-1/4}\right)$, yet it is not guaranteed to be positive-semidefinite. By taking $k_n = \lfloor\theta n^{0.6}\rfloor$, the bias correction term can be ignored and the estimator becomes positive-semidefinite. In this case, the convergence rate reduces to $\mathcal{O}_p\left(n^{-1/5}\right)$.

In practice, the observed log prices are irregularly spaced. We define an information set $\mathcal{F}$ to record the time points with observed transactions.

$$\mathcal{F} = \{t_{ij}|Y_{i,t_{ij}} \text{ is available at } t_j,\ i=1,\ldots,p,\ j=0,\ldots,n\},$$

where t_{ij} represents the time point when the i-th asset is traded at time t_j, that is, the log price $Y_{i,t_{ij}}$ is observable. If $t_{ij} \notin \mathcal{F}$, it means there is no transaction of the ith asset at time t_j. In this case, the corresponding log price Y_{i,t_j} is considered a "missing" value.

3. Main results

We now present the EMP realized-covariance estimator under two scenarios. For synchronous yet noisy data, we extend the Christensen et al. (2010)

MPA estimator by introducing a general eigenvalue correction method in Section 3.1. We show that the correction method ensures positive semidefiniteness without damaging the consistency or the asymptotic limiting distribution of the realized-covariance estimator. For asynchronous and noisy data, we show how to use the HFF technique to generate high-frequency synchronous data by retaining the original cross-dependence. The HFF synchronizing technique and its convergence are detailed in Section 3.2.

3.1 The eigenvalue correction

Suppose an integrated covariance estimator, denoted as S_1, is available. Although Σ is a positive-semidefinite matrix with $\Sigma \geq 0$, or a positive-definite matrix with $\Sigma > 0$, the estimator S_1 may not satisfy the condition of $S_1 \geq 0$ or $S_1 > 0$ due to, for example, a bias correction approach. We propose a general approach to construct a nonnegative-definite estimator S that has the same convergence rate and limiting distribution as the preliminary estimator S_1.

Denote the spectral decompositions of S_1 and Σ by

$$S_1 = U\hat{\Lambda}U^* = \sum_{i=1}^{p} \hat{\lambda}_i \mathbf{u}_i \mathbf{u}_i^*, \quad \Sigma = V\Lambda V^* = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^* \tag{3.1}$$

where $\hat{\lambda}$'s and λ's are the eigenvalues of S_1 and Σ, respectively, and $\mathbf{u}_i$ and $\mathbf{v}_i$ are the orthonormal eigenvectors associated with $\hat{\lambda}_i$ and λ_i for all i. Set all eigenvalues $\hat{\lambda}_i$ to be $|\hat{\lambda}_i|$, $\forall i = 1, \ldots p$. The proposed eigenvalue correction, denoted by S, is performed as:

$$S = U|\hat{\Lambda}|U^*, \tag{3.2}$$

where $|\hat{\Lambda}| = \text{diag}(|\hat{\lambda}_1|, \ldots, |\hat{\lambda}_p|)$. Theorem 1 shows that the estimator S is consistent, if the preliminary estimator S_1 is consistent. The asymptotic distribution of S is derived in Theorem 2, which proves that the estimator has the same asymptotic limiting distribution as S_1.

Theorem 1. *Suppose that $\Sigma \geq 0$ and the maximum eigenvalue of Σ, denoted by λ_{max}, is bounded. Let S_1 be a symmetric matrix satisfying*

$$S_1 - \Sigma \xrightarrow{P} 0. \tag{3.3}$$

Then, S (cf. (3.2)) is a consistent estimator of Σ, that is,

$$S - \Sigma \xrightarrow{P} 0.$$

When reinforcing the condition on Σ, we conclude that S and S_1 have the same limiting distribution as shown in the following theorem.

Theorem 2. *Suppose that $\Sigma > 0$ and $|\lambda_{max}|$ is bounded. Let S_1 be a symmetric matrix satisfying*

$$\alpha_n(S_1 - \Sigma) \xrightarrow{d} Z, \tag{3.4}$$

where $\alpha_n \to \infty$ as $n \to \infty$. Then,

$$\alpha_n(S - \Sigma) \xrightarrow{d} Z.$$

Remark 1. We choose MPA as the preliminary estimator S_1. It has the optimal convergence rate $\mathcal{O}_p(n^{-1/4})$ but is not guaranteed to be positive semidefinite (see Christensen et al., 2010). After the eigenvalue correction procedure, the corrected estimator S has the same optimal convergence rate.

Remark 2. Rebonato and Jäckel (1999) suggested replacing the negative eigenvalues with zeros whereas McNeil et al. (2005) suggested replacing the negative eigenvalues with small positive number. We perform simulations to investigate the numerical performance of our alternative eigenvalue correction approach. We find that the proposed approach improves by 43% and 67% for the relative accuracy of the smallest eigenvalues compared with the methods of Rebonato and Jäckel (1999) and McNeil et al. (2005), respectively. Details can be found in Section S2.1 of the supplementary material.

3.2 Synchronization

For asynchronous data with noise, we present the HFF synchronizing technique to pre-process the data. The role of HFF is to recover/estimate the missing observations in a sequence of synchronous filters obtained by eigen-decomposing the covariance matrix of the lower-frequency sample.

Let S_0 be a covariance estimator of the noisy data. Usually, S_0 is a quadratic variation of the synchronized yet low-frequency data. Due to microstructure noise, the estimator is biased and eventually outputs the sum of the integrated covariance matrix Σ and the microstructure noise variance Ψ. Perform spectral decomposition on S_0:

$$S_0 = \Gamma A \Gamma^\top = \sum_{i=1}^{p} a_i \gamma_i \gamma_i^\top \tag{3.5}$$

where A is a diagonal matrix with eigenvalues a_i on the diagonal and Γ is the matrix of orthonormal eigenvectors.

Assume there exists a linear filter $\boldsymbol{Z}_{t_j}^{(0)} = (Z_{1t_j}^{(0)}, \ldots, Z_{pt_j}^{(0)})^\top$ that is a projection of the unobserved synchronous log returns $\mathbf{R}_{t_j} = \boldsymbol{Y}_{t_j} - \boldsymbol{Y}_{t_{j-1}}$:

$$\mathbf{R}_{t_j} = \Gamma^\top \boldsymbol{Z}_{t_j}^{(0)}, \quad t_j = j/n, \quad j = 1, \ldots, n. \tag{3.6}$$

The linear filter is synchronous and retains the dependence information in the return processes.

Without loss of generality, we assume that the initial value of each asset Y_{i,t_0} exists for $i = 1, \ldots, p$. Denote the synchronous yet noisy log prices $\hat{\mathbf{Y}}_{t_j} = (\hat{Y}_{1t_j}, \ldots, \hat{Y}_{pt_j})^\top$, $j = 0, \ldots, n$. We set $\hat{\mathbf{Y}}_{t_0} = \mathbf{Y}_{t_0}$. Starting from time t_1 onward, the HFF technique iteratively recovers the missing values by minimizing the squared prediction error. At any time t_j, when the previous log prices $\hat{\boldsymbol{Y}}_{t_{j-1}}$ are known, we have the log returns $\hat{R}_{i,t_{ij}} = Y_{i,t_{ij}} - \hat{Y}_{i,t_{j-1}}$. For any $t_{ij} \notin \mathcal{F}$, we evaluate $\hat{Z}_{t_j}$ via the minimizers

$$\begin{aligned} \hat{\boldsymbol{Z}}_{t_j} &= \operatorname{argmin}_{\boldsymbol{Z}_{t_j}} \sum_{i=1}^{p} \left[\left(\hat{R}_{i,t_{ij}} - \hat{\gamma}_i^\top \boldsymbol{Z}_{t_j} \right)^2 I\{t_{ij} \in \mathcal{F}\} \right] \\ &\quad + \delta_n \left(\boldsymbol{Z}_{t_j} + 0.5 \hat{\boldsymbol{Z}}_{t_{j-1}} \right)^\top \hat{A}^{-1} \left(\boldsymbol{Z}_{t_j} + 0.5 \hat{\boldsymbol{Z}}_{t_{j-1}} \right). \end{aligned} \tag{3.7}$$

where the first part of (3.7) focuses on the projection errors. As such, the filtered series are disciplined by the eigendecomposition of S_0 and retain the stability of the cross-dependence structure. Due to asynchrony, an optimization of the prediction error alone does not produce a unique solution. A smoothing penalty, that is, the second part of (3.7), is introduced. It links the estimated filter with the previous (known) values. We standardize the filtered series by its own variance, the eigenvalues of $\hat{A}$. This penalty selection not only ensures the continuity of the filtering procedure but also incorporates the first-order autocorrelation in the noisy data. The tuning parameter δ_n controls the level of smoothness of the filtered series. While large values of δ_n lead to over-smoothing, small values may create an unnecessarily rough process. Cross-validation is used to select the optimal value of δ_n. It turns out that its order is proportional to the inverse of the eigenvalues. Finally, the filtered log price is obtained by $\hat{\boldsymbol{Y}}_{t_j} = \hat{\Gamma}^\top \hat{\boldsymbol{Z}}_{t_j} + \hat{\boldsymbol{Y}}_{t_{j-1}}$ where $\hat{\Gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)$ and $\hat{A} = \mathrm{diag}\{\hat{a}_1, \ldots, \hat{a}_p\}$ are the eigenvectors and eigenvalues of the estimator S_0. If not all of the assets are traded at time t_j, there is no benefit from using the dependence across assets compared with the previous tick technique. We set $\hat{\boldsymbol{Y}}_{t_j} = \hat{\boldsymbol{Y}}_{t_{j-1}}$.

The formal algorithm of the HFF technique is presented as follows: Set $j = 1$. Let $\hat{\boldsymbol{Z}}_{t_0} = \boldsymbol{0}_p$ and we have $\hat{S} = \hat{\Gamma}\hat{A}\hat{\Gamma}^\top$.

1. If $t_{ij} \notin \mathcal{F}$ for all $i = 1, \ldots, p$, set $\hat{\boldsymbol{Y}}_{t_j} = \hat{\boldsymbol{Y}}_{t_{j-1}}$ and jump to step 4.

2. If $t_{ij} \in \mathcal{F}$ with at least one $i = 1, \ldots, p$, compute log return $\hat{R}_{i,t_{ij}} =$

$Y_{i,t_{ij}} - \hat{Y}_{i,t_{j-1}}$ for every i satisfying $t_{ij} \in \mathcal{F}$.

3. Obtain the linear filter $\hat{\boldsymbol{Z}}_{t_j}$ that minimizes the objective function (3.7). We have $\hat{\boldsymbol{Y}}_{t_j} = \hat{\Gamma}^\top \hat{\boldsymbol{Z}}_{t_j} + \hat{\boldsymbol{Y}}_{t_{j-1}}$.

4. Stop until $j = n$; otherwise, renew $j = j + 1$ and return to step 1.

We also investigate the convergence of the proposed filtering technique.

Theorem 3. *Assume that $\hat{S} - S_0 = O_p(n^{-1/4})$. Then, for all $j = 1, 2, \ldots, n$, we have:*

$$\|\hat{\boldsymbol{Z}}_{t_j} - \boldsymbol{Z}^{(0)}_{t_j}\| = O_p(n^{-1/4}) + O(\delta_n) + O(m_j), \tag{3.8}$$

where m_j represents the number of missing values of Y_{i,t_j} at time t_j. Moreover,

$$\frac{1}{n}\sum_{j=1}^{n} \|\hat{\boldsymbol{Z}}_{t_j} - \boldsymbol{Z}^{(0)}_{t_j}\| = O_p(n^{-1/4}),$$

if $\delta_n = O(n^{-1/4})$ and $n^{-1}\sum_{j=1}^{n} m_j = O(n^{-1/4})$.

3.3 The efficient and positive semidefinite pre-averaging estimator

Given the synchronized high-frequency data $\hat{\boldsymbol{Y}}_{t_j}$ from Section 3.2, the pre-averaging estimator is computed and denoted as S_1; see Section 3.1:

$$S_1 = \frac{n}{n - k_n + 2}\frac{12}{k_n}\sum_{j=0}^{n-k_n+1} \bar{\mathbf{Y}}^n_{t_j}(\bar{\mathbf{Y}}^n_{t_j})^\top - \frac{12}{2n\theta^2}\sum_{j=1}^{n}(\hat{\boldsymbol{Y}}_{t_j} - \hat{\boldsymbol{Y}}_{t_{j-1}})(\hat{\boldsymbol{Y}}_{t_j} - \hat{\boldsymbol{Y}}_{t_{j-1}})^\top, \tag{3.9}$$

where $\bar{\mathbf{Y}}^n_{t_j} = \frac{1}{k_n}\left(\sum_{\ell=k_n/2}^{k_n-1}\hat{\mathbf{Y}}_{t_{j+\ell}} - \sum_{\ell=0}^{k_n/2}\hat{\mathbf{Y}}_{t_{j+\ell}}\right)$ and $k_n = \lfloor\theta\sqrt{n}\rfloor$ with a given constant $\theta > 0$. It is unbiased but not guaranteed to be positive-semidefinite. Decompose S_1 as in (3.1) and take the absolute value of the eigenvalues. We obtain the efficient and positive-semidefinite pre-averaging estimator S in (3.2) and name the estimator efficient multiple pre-averaging (EMP). We show that the EMP estimator is consistent with the optimal convergence rate at $\mathcal{O}_p\left(n^{-1/4}\right)$ in Theorem 4 of the Appendix with two additional assumptions.

4. Simulation study

In this section, we run a series of simulations to investigate the performance of the proposed EMP estimator. We compare the EMP estimator with the following two popular alternative estimators:

- MPA: multiple pre-averaging estimator with the synchronizing technique of Hayashi and Yoshida proposed by Christensen et al. (2010) (cf. (2.6));
- MK: kernel estimator with the refresh time synchronization technique proposed by Barndorff-Nielsen et al. (2011).

Moreover, given that MPA showed the best performance in a previous analysis, we investigate the individual effects of the components of the EMP estimator in the MPA framework, namely the proposed eigenvalue correction approach (MPA-E), the HFF technique (MPA-H), and the negative first-order autocorrelation (MPA-N).

- MPA-E: MPA estimator with only the proposed eigenvalue correction;
- MPA-H: MPA estimator with only the HFF technique;
- MPA-N: MPA estimator with the proposed eigenvalue correction and HFF approach but excluding the negative first-order autocorrelation.

In other words, MPA-N is the same as EMP except that the high-frequency filtration is performed by minimizing the following function

$$\hat{\boldsymbol{Z}}_{t_j} = \operatorname{argmin}_{\boldsymbol{Z}_{t_j}} \sum_{i=1}^{p} \left[\left(\hat{R}_{i,t_{ij}} - \hat{\gamma}_i^\top \boldsymbol{Z}_{t_j} \right)^2 I\{t_{ij} \in \mathcal{F}\} \right] + \delta_n \boldsymbol{Z}_{t_j}^\top \hat{A}^{-1} \boldsymbol{Z}_{t_j}.$$

By comparing MPA-N and EMP, we want to see how the HFF improves the estimation of covariance matrices by incorporating the empirical feature of the negative first-order autocorrelation.

We generate noisy and asynchronous processes under various scenarios with dimensions $p = 5,\ 10,$ and 15. The simulation contains three real datasets oriented with parameters learned from Trade and Quote (TAQ) data in the finance, electronics, and food sectors. We also experiment on four extreme scenarios to investigate the performance of the EMP estimator.

4.1 Set up

We first generate efficient and synchronous log prices $\mathbf{X}_t$ of p assets, following the setup in Wang and Zou (2010).

$$d\mathbf{X}_t = \boldsymbol{\sigma}_t^\top d\mathbf{B}_t, \quad t \in [0,1],$$

where $\mathbf{B}_t = (B_{1t}, \ldots, B_{pt})^\top$ is a standard p-dimensional Brownian motion and $\boldsymbol{\sigma}_t$ is a Cholesky decomposition of $\Sigma_t = (\Sigma_{ij,t})_{1\le i,j\le p}$, which is defined below. Let the diagonal elements of Σ_t follow a Cox-Ingersoll-Ross (CIR) process,

$$d\Sigma_{ii,t} = \theta_i(\mu_i - \Sigma_{ii,t})dt + \omega_i\sqrt{\Sigma_{ii,t}}dW_{it},$$

where μ_i denotes the long-term mean of the volatility, $i = 1, \ldots, p$ and W_{it} are standard one-dimensional Brownian motion independent of $\mathbf{B}_t$. Define the off-diagonal elements by

$$\Sigma_{ij,t} = [\kappa(t)]^{|i-j|}\sqrt{\Sigma_{ii,t}\Sigma_{jj,t}}, \quad 1 \le i \ne j \le p,$$

where $\kappa(t)$ is given by

$$\begin{aligned}
\kappa(t) &= \frac{e^{2u(t)}-1}{e^{2u(t)}+1}, \quad du(t) = 0.3[0.64 - u(t)]dt + 0.118u(t)dW_{\kappa,t}, \\
W_{\kappa,t} &= \sqrt{0.96}W^0_{\kappa,t} - 0.2\sum_{i=1}^{p} B_{it}/\sqrt{p},
\end{aligned}$$

and $W^0_{\kappa,t}$ is a standard one-dimensional Brownian motion independent of $\mathbf{B}_t$ and W_{it}.

The synchronous yet noisy log prices are generated with Gaussian noise:

$$\mathbf{Y}_{t_j} = \mathbf{X}_{t_j} + \boldsymbol{\epsilon}_{t_j},$$

where $t_j = j/n$ with $j = 0, \ldots, n$, and $\boldsymbol{\epsilon}$ is an i.i.d. random vector with mean zero and variances η_i, $i = 1, \ldots, p$. Then, the asynchronous and noisy price processes are generated by sampling from Poisson processes with intensity $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^\top$. Note that the generated processes on average have $23\,400/\psi_1$ to $23\,400/\psi_p$ observations.

For parameter settings, we consider three practically oriented experiments based on the TAQ data in the finance, electronics, and food sectors. For each sector, the variances of microstructure noise (η_i), the long-term means of the volatility (μ_i), and the intensities (ψ_i) are respectively estimated from five arbitrarily selected assets; see Table S4 in the supplementary material. For the extreme scenarios, we design four experiments as follows:

- Noisy: a lower signal-to-noise ratio range from 0.017 to 0.034;
- Ex-Asy: dissimilar sampling frequencies with $\psi_i = 3 \sim 60$;
- Ex-HF: ultra-high sampling frequencies with $\psi_i = 3 \sim 5$;
- Negative: an artificial signal-to-noise ratio ranging from 0.00043 to 0.017 for estimating negative-definite covariance matrices.

Detailed parameter settings of each scenario are listed in Table S4. In addition, the parameter settings of $p = 10$ assets are combined with the parameter settings of the finance and electronics sectors, and the parameter settings of $p = 15$ assets are combined with the parameter settings of the finance, electronics, and food sectors. In each sector, the sample size is $n = 23\,400$ with replications $m = 1\,000$ times.

Following initial screening, there are 230 times out of 1 000 replications in which the MPA estimator is not positive-semidefinite for the Negative scenario, and the negative eigenvalues mostly occur for the fifth eigenvalue. For the cases of $p = 10$ and $p = 15$, the frequencies at which each eigenvalue is negative for the 1 000 replications are plotted in Figure 1. Overall, the frequencies of nonpositive-semidefinite covariance estimators are 99% for the cases of both $p = 10$ and $p = 15$.

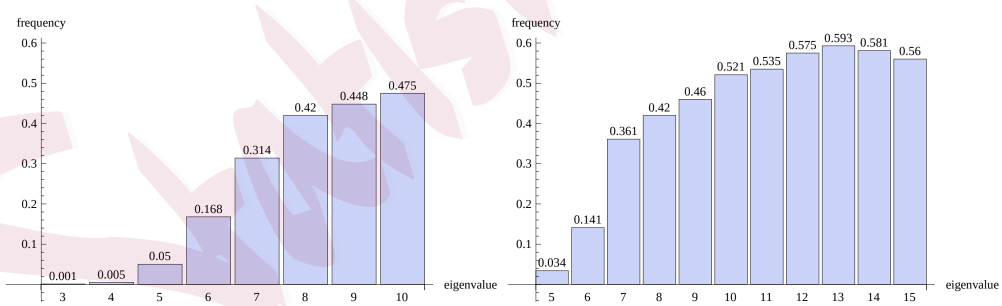

Figure 1: The occurrence frequencies of negative eigenvalues based on $1,000$ replications for $p = 10$ (left panel) and $p = 15$ (right panel).

For each scenario, the EMP estimator is obtained by

1. filtering the high-frequency synchronous data by the HFF technique, in which the tuning parameter δ_n is chosen by cross-validation.

2. perform proposed eigenvalue correction and obtain the realized-covariance estimator.

4.2 Evaluation and alternatives

We measure both the overall and element-wise accuracy of the EMP estimator. The overall performance is evaluated by the relative error (RE) of each eigenvalue, defined as

$$RE_i = \frac{\sqrt{\frac{1}{m}\left[\sum_{s=1}^{m}(\hat{\lambda}_i^{(s)} - \lambda_i)^2\right]}}{\lambda_i}, \quad i = 1, \ldots, p,$$

where $\lambda_i/\hat{\lambda}_i^{(s)}$ denotes the i-th true/estimated eigenvalues of the s-th replication. The maximum norm (MN) evaluates the element-wise accuracy measured by the largest absolute deviation of all elements.

$$MN = \frac{1}{m}\sum_{s=1}^{m}\left\{\max_{i,j}|\widehat{\Sigma}_{ij}^{(s)} - \Sigma_{ij}|\right\},$$

where Σ_{ij} is the (i,j)-th element of the covariance and $\hat{\Sigma}_{ij}^{(s)}$ is the estimated (i,j)-th element in the s-th replication, $i,j = 1,\cdots,p$, $s = 1,\ldots,m = 1\ 000$. The lower of these two measures represents the better accuracy of the estimated covariance matrix.

Table 1 reports the RE and MN of the EMP estimator. It shows that the EMP estimator provides accurate results with low estimation errors in the finance, electronics, food, Noisy, Ex-Asy, and Ex-HF sectors. The MPA and MK alternative estimators are compared with the EMP estimator by calculating the ratios of their errors to the corresponding EMP measurements. A ratio larger than 1 indicates a poorer accuracy of the alternative. In the real data-oriented scenarios (finance, electronics, and food), the EMP estimator, although far from optimal, still has lower relative errors compared with MPA and MK. More specifically, the improvement in relative overall performance ranges from 1.7% (RE_4 electronics) to 77% (RE_5 finance) and the improvement in element-wise accuracy ranges from 60.8% (finance) to 66.2% (electronics) compared with the MPA. Compared with MK, with the exception of the first eigenvalue in finance, the EMP estimator displays an increase in overall accuracy ranging from 8.4% (RE_3 finance) to 198.1% (RE_1 food) and an improvement of more than 246.2% (finance) in terms of MN. In the extreme scenarios (Noisy, Ex-Asy, and Ex-HF), the EMP estimator outperforms MK and MPA without exception. The EMP estimator enhances the element-wise accuracy by a range of 69.5% (compared with MPA Noisy) to 311.7% (compared with MK Ex-HF).

Table 2 presents the REs of MPA, MPA-E, MPA-H, MPA-N, and EMP in the Negative scenario. The results confirm that the proposed eigenvalue correction approach can efficiently improve the negative eigenvalue(s) by around 20.4%. Error correction contributes greatly in the cases with smaller eigenvalues, especially those close to zero or negative. For the larger eigenvalues, there is little benefit of using the error correction approach. The HFF technique, conversely, leads to a big improvement in the larger eigenvalues by utilizing the cross-dependence among the multiple assets. However, HFF does not provide significant benefits for the smaller eigenvalues, which represent fewer features of the covariance matrix. Meanwhile, the MPA-N results indicate the importance of incorporating the negative autocorrelation in the HFF technique. Without these empirical features, MPA-N produces a mixture of beneficial and detrimental contributions but an overall decrease in accuracy. The EMP estimator nicely combines the two techniques and the empirical features. It benefits in the larger-eigenvalue cases from the richer information of multiple assets and in the smaller-eigenvalue cases from the correction of negative values.

The results of $p = 10$ and $p = 15$ are similar and we discuss this in Section S2.3 of the supplementary material.

In summary, the EMP estimator displays substantial improvements in relative performance and is stable in different scenarios, indicating that it estimates the true covariance matrix with reasonable accuracy.

5. Real data analysis

In this section, we implement the synchronizing technique and eigenvalue correction approach to apply the EMP realized-covariance estimator to real-world tick-by-tick financial data. We also apply the proposed EMP esti-

Table 1: Comparisons of EMP and alternative realized covariance estimators in terms of the RE of eigenvalues and maximum norms (MN).

	Food			Electronics			Finance		
	EMP	$\frac{\text{MK}}{\text{EMP}}$	$\frac{\text{MPA}}{\text{EMP}}$	EMP	$\frac{\text{MK}}{\text{EMP}}$	$\frac{\text{MPA}}{\text{EMP}}$	EMP	$\frac{\text{MK}}{\text{EMP}}$	$\frac{\text{MPA}}{\text{EMP}}$
RE_1	0.162	2.981	1.265	0.170	2.747	1.076	0.218	0.982	1.023
RE_2	0.142	2.951	1.577	0.145	2.725	1.539	0.159	1.390	1.226
RE_3	0.156	2.365	1.455	0.143	2.839	1.434	0.215	1.084	1.121
RE_4	0.183	2.891	1.426	0.230	1.913	1.017	0.200	1.740	1.165
RE_5	0.272	2.419	1.232	0.265	2.525	1.260	0.236	2.436	1.771
MNs	$6.98E-5$	3.734	1.643	$6.72E-5$	3.853	1.662	$7.02E-5$	3.462	1.608
	Noisy			Ex-Asy			Ex-HF		
	EMP	$\frac{\text{MK}}{\text{EMP}}$	$\frac{\text{MPA}}{\text{EMP}}$	EMP	$\frac{\text{MK}}{\text{EMP}}$	$\frac{\text{MPA}}{\text{EMP}}$	EMP	$\frac{\text{MK}}{\text{EMP}}$	$\frac{\text{MPA}}{\text{EMP}}$
RE_1	0.176	2.892	1.250	0.209	2.909	1.038	0.138	2.928	1.326
RE_2	0.165	2.679	1.479	0.172	2.831	1.767	0.109	3.367	1.725
RE_3	0.341	1.067	0.633	0.217	2.037	1.258	0.256	1.457	0.738
RE_4	0.253	1.767	0.874	0.222	2.824	1.500	0.143	2.832	1.294
RE_5	0.303	2.205	1.218	0.339	2.469	2.555	0.168	3.595	1.595
MNs	$5.11E-5$	3.560	1.695	$9.08E-5$	3.992	1.800	$9.13E-5$	4.117	1.780

Table 2: Comparison of MPA, MPA-E, MPA-H, MPA-N and EMP in terms of REs.

	RE_1	RE_2	RE_3	RE_4	RE_5
MPA	0.4688	0.3702	0.5086	0.5799	0.8648
MPA-E	0.4688	0.3702	0.5086	0.5479	0.7182
MPA-H	0.2711	0.2607	0.2616	0.2258	1.4661
MPA-N	0.3003	0.8231	0.6509	0.5761	1.6064
EMP	0.2711	0.2607	0.2616	0.2258	0.7647

mator in portfolio selection to show its usefulness in financial applications, where covariance is a key input factor.

We consider the TAQ data of seven assets listed on the New York Stock Exchange (NYSE): AIG, GE, IBM, JPM, MRK, PFE, and T, over the period 2005/01/02∼2005/12/31. The normal trading hours of the NYSE are from 9:30 to 16:00, or 6.5 hours (23,400 seconds). We remove the 7 days from 2005/11/21∼2005/11/30 due to the unavailable data of asset T. In total, there are 245 trading days. For illustration, Figure 2 depicts the evolution of the daily adjusted closing prices of the PFE and MRK stocks in 2015. The two assets belong to the same industry, that is, pharmaceuticals, and hence are naturally positively correlated; the historical correlation estimator is 0.51.

The two alternative estimators MPA and MK are also computed based on the high-frequency data. Figure 3 depicts the time plot of the estimated daily correlations of PFE and MRK. Each of the realized-covariance estimators delivers positive correlations in most cases, with values varying in the range of $[-0.53, 0.87]$ for EMP, a larger range of $[-0.74, 0.98]$ for MPA, and a range of $[-0.31, 0.98]$ for MK. The correlation between PFE and MRK becomes negative after day 200, which is well represented by the EMP estimator but not by the alternatives, an observation that supports the general accuracy of the EMP estimator. Furthermore, even with a large $n = 12,552$, there are 40 days when the MPA estimators are not positive-semidefinite.

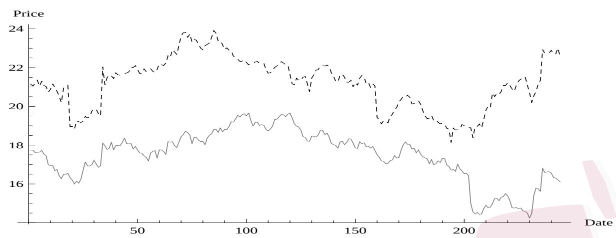

Figure 2: Time plot of daily closing prices of PFE (gray real line) and MRK (black dashed line) stocks for a total of 245 days.

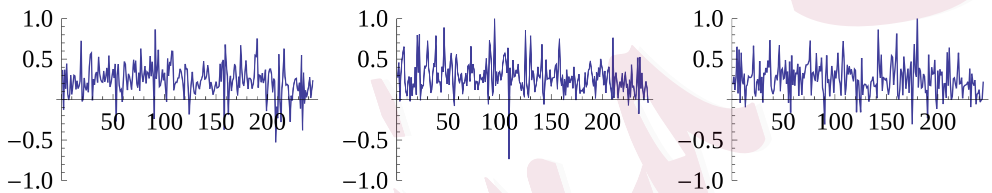

Figure 3: Time plots of the intra-daily correlation estimations between PFE and MRK stocks based on EMP (left panel), MPA (middle panel) and MK (right panel), respectively, for a total of 245 days.

5.1 Application in portfolio allocation

Markowitz mean-variance portfolio selection has a profound impact on financial economics. Suppose that $\mathbf{w}$ represent the weights of portfolio allocation with the constraint of $\mathbf{w}^*\mathbf{1} = 1$. The Markowitz mean-variance optimization is equivalent to maximizing the following function

$$M(\boldsymbol{\mu}, \Sigma) = \mathbf{w}^*\boldsymbol{\mu} - \lambda \mathbf{w}^*\Sigma\mathbf{w}$$

which is sensitive to estimation errors in the expected return and the covariance matrix, especially when the portfolio is large. Fan et al. (2012) showed that the estimation errors can be bounded as

$$|M(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) - M(\boldsymbol{\mu}, \Sigma)| \leq \|\hat{\boldsymbol{\mu}} - \mu\|_\infty \|\mathbf{w}\|_1 + \lambda \|\hat{\Sigma} - \Sigma\|_\infty \|\mathbf{w}\|_1,$$

where $\|\cdot\|_\infty$ refers to the maximum component-wise estimation errors. The problem disappears when the gross-exposure constraint $\|\mathbf{w}\|_1 \leq c$ is imposed for a moderate c, where c is the total exposure allowed:

$$\min_{\mathbf{w}} \mathbf{w}^* \Sigma \mathbf{w} \qquad s.t. \ \|\mathbf{w}\|_1 \leq c \text{ and } \mathbf{w}^* \mathbf{1} = 1.$$

Let $R(\mathbf{w}, \Sigma) = \mathbf{w}^* \Sigma \mathbf{w}$, Fan et al. (2012) showed that

$$|R(\mathbf{w}, \hat{\Sigma}) - R(\mathbf{w}, \Sigma)| \leq \|\hat{\Sigma} - \Sigma\|_\infty \|\mathbf{w}\|_1.$$

The above estimation errors do not accumulate in the risk. Fan et al. (2012) extended the work of Fan et al. (2012) to high-frequency data by using the two-scale realized-covariance estimator combined with the all-refresh and pairwise-refresh synchronizing techniques.

Following the above work, we construct portfolios based on tick-by-tick records. The optimal weights are updated by the realized-covariance estimator from the previous day:

$$\min_{\mathbf{w}} \mathbf{w}^* \Sigma \mathbf{w} \qquad s.t. \ \|\mathbf{w}\|_1 \leq c \text{ and } \mathbf{w}^* \mathbf{1} = 1.$$

where we consider three cases with $c = 1$, $c = 2$, and $c = 3$ using three alternative realized-covariance estimators, MPA, MK, and EMP.

Table 3 reports the statistical summary of the portfolios based on different realized-covariance estimators and $c = 1, 2$, and 3. Without exception, the EMP portfolios are always better than the MPA and MK specifications. The EMP portfolio is the only portfolio with a positive mean, and it produces the smallest standard deviation. In most cases, the EMP portfolio outperforms the alternatives when extreme loss is considered and $c = 2$ and 3. The other portfolios for extreme losses are also competitive with the best solutions. To visualize the differences between estimators, we plot the histograms of the daily log returns of the portfolios with different realized-covariance estimators in Figure 4. Moreover, the EMP portfolio provides superior performance regarding cumulative returns, as displayed in Figure 5 with $c = 1$, $c = 2$, and $c = 3$. After $t = 130$, the EMP portfolio outperforms the alternatives and the equal-weighted portfolio. Figure 5 also depicts the daily portfolio volatility $(\hat{\mathbf{w}}^*_{\text{opt}}\hat{\Sigma}\hat{\mathbf{w}}_{\text{opt}})$ using different realized-covariance estimators. The results, as summarized in Table 4, indicate that the EMP portfolio has a greater chance of obtaining lower portfolio volatilities compared with MPA and MK. To summarize, the EMP estimator is superior in the Markowitz mean-variance portfolio selection experiment.

6. Conclusion and future work

We have developed a new realized-covariance estimator that simultaneously ensures positive-semidefiniteness and optimal efficiency. By drawing on the

Table 3: The medians, means, standard deviations (Std.), and 1% and 5% quantiles of the log returns of portfolio prices based on three covariance matrix estimators.

	$c = 1$				
	Median	Mean	Std.	1% quantile	5% quantile
EMP	$1.40E-4$	$4.67E-5$	$6.87E-3$	$-1.12E-2$	$-1.57E-2$
MK	$0.92E-4$	$-1.93E-5$	$7.42E-3$	$-1.17E-2$	$-1.65E-2$
MPA	$-1.24E-4$	$-1.46E-4$	$6.94E-3$	$-1.08E-2$	$-1.53E-2$
	$c = 2$				
	Median	Mean	Std.	1% quantile	5% quantile
EMP	$1.39E-4$	$1.89E-5$	$6.93E-3$	$-1.12E-2$	$-1.57E-2$
MK	$-0.21E-4$	$-8.09E-5$	$8.85E-3$	$-1.17E-2$	$-2.08E-2$
MPA	$-0.96E-4$	$-9.72E-5$	$7.21E-3$	$-1.08E-2$	$-1.59E-2$
	$c = 3$				
	Median	Mean	Std.	1% quantile	5% quantile
EMP	$1.40E-4$	$2.56E-5$	$6.93E-2$	$-1.12E-2$	$-1.56E-2$
MK	$-0.58E-4$	$-1.21E-4$	$9.37E-2$	$-1.29E-2$	$-2.05E-2$
MPA	$-1.24E-4$	$-9.84E-5$	$7.38E-2$	$-1.12E-2$	$-1.96E-2$

Table 4: Portions of the smallest portfolio volatility of different covariance estimators.

	$c = 1$	$c = 2$	$c = 3$
EMP	44.64%	50%	52.57%
MPA	30.90%	31.62%	31.62%
MK	24.46%	18.38%	15.81%

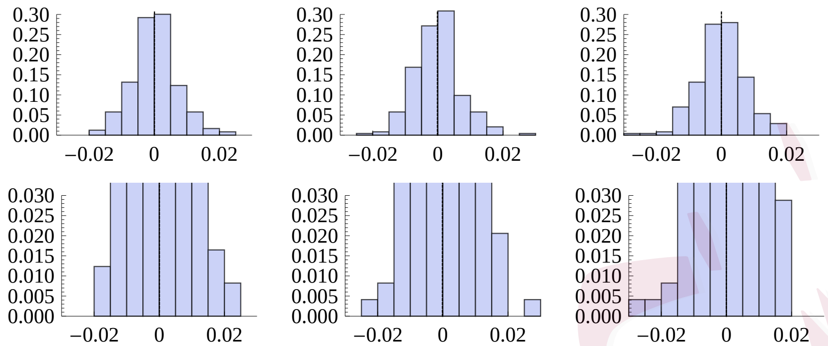

Figure 4: Histograms of the log returns of portfolio prices based on EMP (left panel), MPA (middle panel), and MK (right panel). The lower panel is zoomed in to show the tail sections.

dependence information of data, we were able to iteratively synchronize asynchronous high-frequency data. Together with a correction approach, the developed estimator is positive-semidefinite and efficient at the optimal convergence rate $\mathcal{O}_p\left(n^{-1/4}\right)$. It is consistent and has the same limiting distribution as the efficient estimator. Real data-oriented simulation experiments demonstrated the finite-sample performance of the estimator. Compared with several alternatives, the efficient and positive-semidefinite estimator provided the best accuracy in various experiments. Real data analysis illustrated the superior performance of the proposed estimator in portfolio allocations. The proposed methods are general and could be further applicable to other realized measures and to matrix corrections. In this study, we considered the multidimensional covariance matrix estimation. The extension to high- and large-dimensional covariance matrix estimation is of practical interest, and we leave this to future study. Some important works in this context include but are not limited to Aït-Sahalia and Xiu (2017), Dai et al. (2017), Fan et al. (2016), Kim et al. (2018), Kim et al. (2016), Kong (2017), and Kong (2018).

Supplementary Materials

All of the proofs and some simulations can be found in the supplementary material.

Acknowledgements The first author acknowledges that this work was supported by Ministry of Science and Technology, Taiwan, under grant MOST 105-2628-M-006-001-MY3. Chen gratefully acknowledges the financial support of Singapore Ministry of Education Academic Research Fund Tier 1 at National University of Singapore. G.M. Pan's research has been partially financially supported by Tier 1 grant: RG133/18 and Tier 2 grant: MOE2018-T2-2-112 at NTU both from Ministry of Education, Singapore. This work was supported by the Russian Science Foundation (RSF) grant 19-71-30020 and by the Excellence Cluster Math+ Berlin, project AA4-2.

References

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

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

Andersen, T. G. and T. Bollerslev (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review 39*, 885–905.

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

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

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

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

Barndorff-Nielsen, O. E. and N. Shephard (2002a). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society B 64*, 253–280.

Barndorff-Nielsen, O. E. and N. Shephard (2002b). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics 17*, 457–477.

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

Christensen, K., M. Podolskij, and M. Vetter (2013). On covariance estimation for multivariate continuous iˆto semimartingales with noise in non-synchronous observation schemes.

Journal of Multivariate Analysis 120, 59–84.

Corsi, F., F. Peluso, and F. Audrino (2015). Missing in asynchronicity: A Kalman-EM approach for multivariate realized covariance estimation. *Journal of Applied Econometrics 30*, 377–397.

Dacorogna, M., R. Gençay, U. Müller, R. Olsen, and O. Pictet (2001). *An introduction to high-frequency finance*. Academic Press New York.

Dai, C., K. Lu, and D. Xiu (2017). Knowing factors or factor loadings, or neither? evaluating estimators of large covariance matrices with noisy and asynchronous data. working paper.

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

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

Fan, J., J. Zhang, and K. Yu (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association 107*(498), 592–606.

French, K. R., G. W. Schwert, and R. F. Stambaugh (1987). Expected stock returns and volatility. *Journal of Financial Economics 19*, 3–30.

Hansen, P. R. and A. Lunde (2006). Realized variance and market microstructure noise. *Journal of Business and Economic Statistics 24*, 127–161.

Hautsch, N., L. M. Kyj, and R. C. Oomen (2012). A blocking and regularization approach to high dimensional realized covariance estimation. *Journal of Applied Econometrics 27*, 625–645.

Ikeda, S. S. (2016). A bias-corrected estimator of the covariation matrix of multiple security prices when both microstructure effects and sampling durations are persistent and endogenous. *Journal of Econometrics 193*, 203–214.

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

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

Kim, D., Y. Wang, and J. Zou (2016). Asymptotic theory for large volatility matrix estimation based on high-frequency financial data. *Stochastic Processes and their Applications 126*, 3527–3577.

Kong, X.-B. (2017). On the number of common factors with high-frequency data. *Biometrika 104*(2), 397–410.

Kong, X.-B. (2018). On the systematic and idiosyncratic volatility with large panel high-frequency data. *Annals of Statistics*. forthcoming.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton University Press: Princeton and Oxford.

Rebonato, R. and P. Jäckel (1999). The most general methodology to create a valid correlation matrix for risk management and option pricing purposes. QUARC preprint.

Schaeffer, L. R. (2014). Making covariance matrices positive definite. working paper.

Shephard, N. and D. Xiu (2017). Econometric analysis of multivariate realised qml: Estimation of the covariation of equity prices under asynchronous trading. *Journal of Econometrics 201*, 19–42.

Varneskov, R. T. (2015). Flat-top realized kernel estimation of quadratic covariation with non-synchronous and noisy asset prices. *Journal of Business and Economic Statistics*. forthcoming.

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

Wasserfallen, W. and H. Zimmermann (1985). The behavior of intra-daily exchange rates. *Journal of Banking and Finance 9*(1), 55–72.

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

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

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

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business and Economic Statistics 14*, 45–52.

Department of Statistics, National Cheng Kung University, Tainan, Taiwan.

E-mail: lclin@mail.ncku.edu.tw

Department of Mathematics, National University of Singapore and Risk Management Institute, National University of Singapore, Singapore.

E-mail: mathcheny@nus.edu.sg

School of Physical & Mathematical Sciences, Nanyang Technological University, Singapore.

E-mail: GMPAN@ntu.edu.sg

WIAS and HU Berlin, HSE and IITP RAS, Moscow.

E-mail: spokoiny@wias-berlin.de

Appendix

Theorem 4. *Let S_{true} be the multiple pre-averaging estimation based on efficient but unobservable log prices.*

$$S_{true} = \frac{n}{n-k_n+2}\frac{12}{k_n}\sum_{j=0}^{n-k_n+1}\bar{\mathbf{Y}}_{t_j}^{n,(0)}(\bar{\mathbf{Y}}_{t_j}^{n,(0)})^\top - \frac{12}{2n\theta^2}\sum_{j=1}^{n}(\boldsymbol{Y}_{t_j}-\boldsymbol{Y}_{t_{j-1}})(\boldsymbol{Y}_{t_j}-\boldsymbol{Y}_{t_{j-1}}),^\top$$

where $\bar{\mathbf{Y}}_{t_j}^{n,(0)} = \frac{1}{k_n}\left(\sum_{\ell=k_n/2}^{k_n-1}\boldsymbol{Y}_{t_{j+\ell}} - \sum_{\ell=0}^{k_n/2}\boldsymbol{Y}_{t_{j+\ell}}\right)$. The asymptotic distribution of S_{true} is given in Christensen et al. (2010) with convergence rate $n^{1/4}$. Let the assumptions of Theorem 2 and 3 hold, and further assume that

(i) $\hat{\boldsymbol{X}}_{t_j} - \boldsymbol{X}_{t_j}$ *and* $\hat{\boldsymbol{\varepsilon}}_{t_j} - \boldsymbol{\varepsilon}_{t_j}$ *share the same order with* $\hat{\mathbf{R}}_{t_j} - \mathbf{R}_{t_j}$*;*

(ii) $\hat{\boldsymbol{\varepsilon}}_{t_j}$ *have lower dependency on each other,* $j = 1, \ldots, n$*.*

We have

$$\|n^{1/4}(S_1 - S_{true})\| = o_p(1),$$

which implies that S_1 has the same limiting distribution as S_{true}.

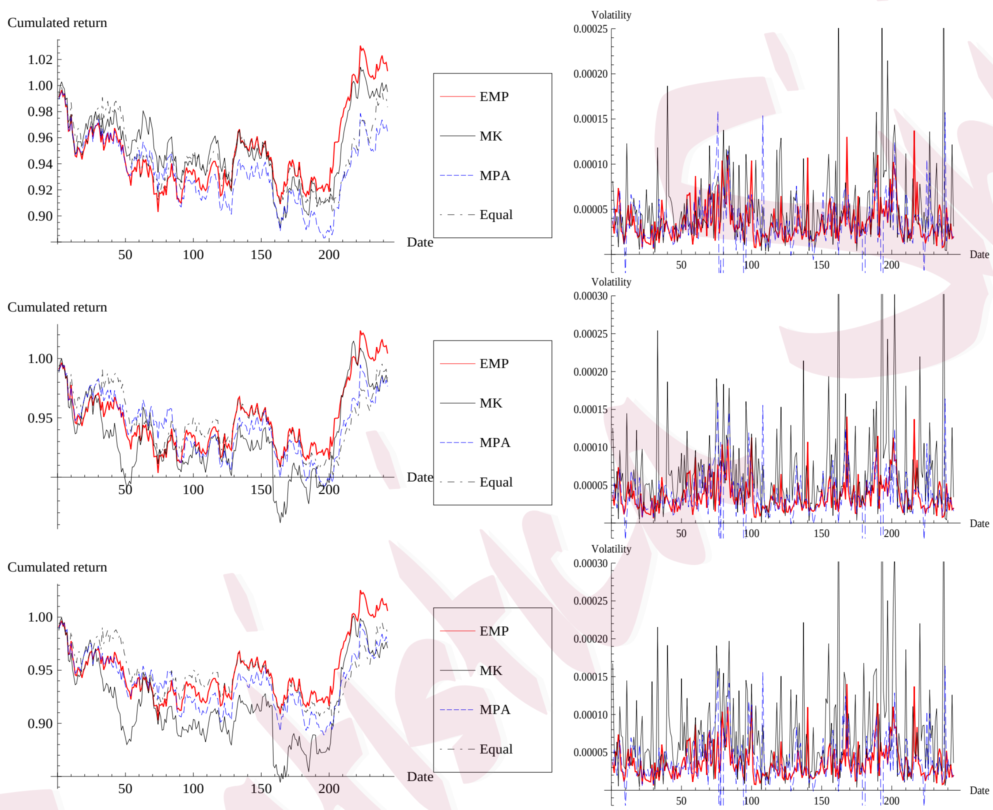

Figure 5: Cumulative portfolio returns (left panel) and portfolio volatilities (right panel) based on EMP (thick solid line), MK (solid line), MPA (dashed line), and equal weights (dot-dash line) for $c = 1$ (upper panel), $c = 2$ (middle panel), and $c = 3$ (lower panel).