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*

Supplementary Material

Section S1 contains all of the proofs, including Theorem 1, 2, 3, and 4.

Some relevant simulations are shown in Section S2.

S1 Proofs

S1.1 Proof of Theorem 1

Before proving the following theorems, we cite the following lemma from Bai et al. (1991) for readers' convenience.

Lemma 1. *Bai et al. (1991)*

Let $A = (a_{ik})$ and $B = (b_{ik})$ be two Hermitian $p \times p$ matrices with spectral *decompositions*

$$A = \sum_{i=1}^{p} \xi_i \mathbf{u}_i \mathbf{u}_i^*, \ \xi_1 \geq \xi_2 \geq \cdots \geq \xi_p,$$

and

$$B = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^*, \ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p,$$

where ξ*'s and* λ*'s are eigenvalues of* A *and* B*, respectively,* $\mathbf{u}$*'s and* $\mathbf{v}$*'s are orthonormal eigenvectors associated with* ξ*'s and* λ*'s, respectively. Further, we assume that*

$$\lambda_{n_{b-1}+1} = \cdots = \lambda_{n_b} = \tilde{\lambda}_b, \ n_0 = 0 < n_1 < \ldots < n_r = p, \ b = 1, \ldots, r,$$
$$\tilde{\lambda}_1 > \tilde{\lambda}_2 > \cdots > \tilde{\lambda}_r,$$

and that

$$|a_{ik} - b_{ik}| < \alpha, \quad i, k = 1, \ldots, p.$$

Then there is a constant M *independent of* α*, such that*

(a) $|\xi_i - \lambda_i| < M\alpha, \quad i = 1, \ldots, p.$

(b) $$\sum_{i=n_{b-1}+1}^{n_b} \mathbf{u}_i \mathbf{u}_i^* = \sum_{i=n_{b-1}+1}^{n_b} \mathbf{v}_i \mathbf{v}_i^* + \mathbf{C}^{(b)},$$
$$\mathbf{C}^{(b)} = (C_{ik}^{(b)}), \ |C_{ik}^{(b)}| < M\alpha, \ l, k = 1, \ldots, p, \ b = 1, \ldots, r. \qquad \square$$

(proof of Theorem 1) By (3.3),

$$(S_1^2 - \Sigma^2) \;=\; (S_1 - \Sigma)^2 + (S_1 - \Sigma)\Sigma + \Sigma(S_1 - \Sigma), \qquad \text{(S1.1)}$$

we obtain that $S_1^2 - \Sigma^2$ converges to zero in probability. Assume that there are r different values of $(\lambda_1, \ldots, \lambda_p)$, say

$$\lambda_{n_{b-1}+1} = \cdots = \lambda_{n_b} = \tilde{\lambda}_b, \; n_0 = 0 < n_1 < \ldots < n_s = p, \; b = 1, \ldots, r.$$

By Lemma 1 and (S1.1), we have for any $\epsilon > 0$,

$$|\hat{\lambda}_i^2 - \lambda_i^2| < M\epsilon, \; i = 1, \ldots, p, \; \textit{in probability} \tag{S1.2}$$

$$\sum_{i=n_{b-1}+1}^{n_b} \mathbf{u}_i \mathbf{u}_i^* - \sum_{i=n_{b-1}+1}^{n_b} \mathbf{v}_i \mathbf{v}_i^* = \mathbf{C}^{(b)} \tag{S1.3}$$

where $C^{(b)} = (C_{lk}^{(b)})$ with $|C_{lk}^{(b)}| \leq M\epsilon$ *in probability* for all $l, k = 1, \ldots, p$ and $b = 1, \ldots, r$.

It follows from (S1.2), (S1.3) and Slutsky's Theorem that

$$\begin{aligned} S - \Sigma &= \sum_{i=1}^{p} \left(|\hat{\lambda}_i| \mathbf{u}_i \mathbf{u}_i^* - \lambda_i \mathbf{v}_i \mathbf{v}_i^* \right) = \sum_{b=1}^{r} |\hat{\lambda}_b| \sum_{i=n_{b-1}+1}^{n_b} \mathbf{u}_i \mathbf{u}_i^* - \sum_{b=1}^{r} \lambda_b \sum_{i=n_{b-1}+1}^{n_b} \mathbf{v}_i \mathbf{v}_i^* \\ &= \sum_{b=1}^{r} \left(|\hat{\lambda}_b| - \lambda_b \right) \sum_{i=n_{b-1}+1}^{n_b} (\mathbf{u}_i \mathbf{u}_i^*) + \sum_{b=1}^{r} \lambda_b \sum_{i=n_{b-1}+1}^{n_b} (\mathbf{u}_i \mathbf{u}_i^* - \mathbf{v}_i \mathbf{v}_i^*) \\ &\leq M'\epsilon \;\; \textit{in probability}, \end{aligned} \tag{S1.4}$$

which implies $S - \Sigma \xrightarrow{P} 0$. $\square$

S1.2 Proof of Theorem 2

By (3.4) and Slutsky's Theorem, we have

$$\begin{aligned}\alpha_n(S_1^2-\Sigma^2) &= \alpha_n(S_1-\Sigma)^*(S_1-\Sigma)+\alpha_n(S_1-\Sigma)^*\Sigma+\alpha_n\Sigma(S_1-\Sigma)\\ &= \alpha_n^{-1}\left[\alpha_n(S_1-\Sigma)\right]^*\left[\alpha_n(S_1-\Sigma)\right]\\ &\quad+\left[\alpha_n(S_1-\Sigma)\right]^*\Sigma+\Sigma\left[\alpha_n(S_1-\Sigma)\right]\\ &\xrightarrow{d} Z\Sigma+\Sigma Z. \end{aligned} \tag{S1.5}$$

This implies $S_1^2-\Sigma^2=\mathcal{O}_p(\alpha_n^{-1})$. By Lemma 1 we then have

$$|\hat{\lambda}_i^2-\lambda_i^2|<M\alpha_n^{-1},\ i=1,\ldots,p,\ \textit{in probability} \tag{S1.6}$$

$$\sum_{i=n_{b-1}+1}^{n_b}\mathbf{u}_i\mathbf{u}_i^*-\sum_{i=n_{b-1}+1}^{n_b}\mathbf{v}_i\mathbf{v}_i^*=\mathbf{C}^{(b)} \tag{S1.7}$$

where $C^{(b)}=(C_{lk}^{(b)})$ with $|C_{lk}^{(b)}|\le M\alpha_n^{-1}$ *in probability* for all $l,k=1,\ldots,p$ and $b=1,\ldots,r$. Then, by (S1.6), (S1.7), Slutsky's Theorem and an argument similar to (S1.4), we have

$$S-\Sigma\le M'\alpha_n^{-1}\ \textit{in probability}, \tag{S1.8}$$

where M' is some positive constant independent of n.

The condition (S1.8) implies that $\alpha_n(S-\Sigma)$ is tight. Consider a subsequence n_k on which $\alpha_{n_k}(S-\Sigma)$ converges in distribution to a random variable, say Y (here and below, to save notation we still use $S-\Sigma$ rather than their expressions on the subsequence). Therefore, by Slutsky's Theorem, it holds that

$$\alpha_{n_k}(S^2 - \Sigma^2) = \alpha_{n_k}(S - \Sigma)S + \alpha_{n_k}\Sigma(S - \Sigma) \xrightarrow{d} Y\Sigma + \Sigma Y. \tag{S1.9}$$

Evidently, (S1.9) is equivalent to

$$\text{vec}(\alpha_{n_k}(S^2 - \Sigma^2)) \xrightarrow{d} (I \otimes \Sigma + \Sigma \otimes I)\text{vec}(Y), \tag{S1.10}$$

where $\otimes$ is the Kronecker product and vec(Y) is the vectorization of a matrix.

Moreover, (S1.5) can be also rewritten as

$$\text{vec}(\alpha_{n_k}(S_1^2 - \Sigma^2)) = \text{vec}(\alpha_{n_k}(S^2 - \Sigma^2)) \xrightarrow{d} (I \otimes \Sigma + \Sigma \otimes I)\text{vec}(Z). \tag{S1.11}$$

Note that the eigenvalues of $(I \otimes \Sigma + \Sigma \otimes I)$ are $\{\lambda_i + \lambda_j,\ i, j = 1, \ldots, p\}$. The hypothesis $\Sigma > 0$ implies that $(I \otimes \Sigma + \Sigma \otimes I)$ is invertible. Therefore, (S1.10) and (S1.11) ensure that vec$(Z) \stackrel{d}{=}$ vec(Y) and hence $Z \stackrel{d}{=} Y$. This means that Y is unique. This, together with the tightness of $\alpha_n(S - \Sigma)$ (which implies that $\alpha_n(S - \Sigma)$ is relatively compact), implies that $\alpha_n(S - \Sigma) \xrightarrow{d} Z$ which completes the proofs. □

S1.3 Proof of Theorem 3

Since $\hat{S} - S_0 = O_p(n^{-1/4})$, by Lemma 1 we have

$$|\hat{a}_i^2 - a_i^2| < Mn^{-1/4},\ i = 1, \ldots, p,\ \textit{in probability} \tag{S1.12}$$

$$\sum_{i=n_{b-1}+1}^{n_b} \hat{\gamma}_i \hat{\gamma}_i^* - \sum_{i=n_{b-1}+1}^{n_b} \gamma_i \gamma_i^* = \mathbf{C}^{(b)} \tag{S1.13}$$

where $C^{(b)} = (C_{lk}^{(b)})$ with $|C_{lk}^{(b)}| \le Mn^{-1/4}$ *in probability* for all $l, k = 1, \ldots, p$ and $b = 1, \ldots, r$.

Define

$$f(\boldsymbol{Z}_{t_j}) = \sum_{i=1}^{p} \left(\hat{R}_{i,t_{ij}} - \hat{\gamma}_i^\top \boldsymbol{Z}_{t_j} \right)^2 I_{ij} + \delta_n \left(\boldsymbol{Z}_{t_j} + 0.5 \hat{\boldsymbol{Z}}_{t_{j-1}} \right)^\top \hat{A}^{-1} \left(\boldsymbol{Z}_{t_j} + 0.5 \hat{\boldsymbol{Z}}_{t_{j-1}} \right),$$

where $I_{ij} = I\{t_{ij} \in \mathcal{F}\}$. Note that $f(\boldsymbol{Z})$ is continuous and strictly convex. Its first derivative is

$$f'(\boldsymbol{Z}_{t_j}) = -2 \sum_{i=1}^{p} \left(\hat{R}_{i,t_{ij}} - \hat{\gamma}_i^\top \boldsymbol{Z}_{t_j} \right) \hat{\gamma}_i^\top I_{ij} + 2\delta_n \hat{A}^{-1} \left(\boldsymbol{Z}_{t_j} + 0.5 \hat{\boldsymbol{Z}}_{t_{j-1}} \right).$$

Setting $f'(\boldsymbol{Z}_{t_j}) = \mathbf{0}$ and solving respective to $\boldsymbol{Z}_{t_j}$, we obtain

$$\hat{\boldsymbol{Z}}_{t_j} = B^{-1} \left(\sum_{i=1}^{p} \hat{R}_{i,t_{ij}} \hat{\gamma}_i I_{ij} - 0.5 \delta_n \hat{A}^{-1} \hat{\boldsymbol{Z}}_{t_{j-1}} \right),$$

where

$$B = \sum_{i=1}^{p} \hat{\gamma}_i \hat{\gamma}_i^\top I_{ij} + \delta_n \hat{A}^{-1}.$$

Recall (3.6), we have $\boldsymbol{Z}_{t_j}^{(0)} = \Gamma \mathbf{R}_{t_j}$. It follows that

$$\hat{\boldsymbol{Z}}_{t_j} - \boldsymbol{Z}_{t_j}^{(0)} = B^{-1} \left(\sum_{i=1}^{p} \hat{R}_{i,t_{ij}} \hat{\gamma}_i I_{ij} - 0.5 \delta_n \hat{A}^{-1} \hat{\boldsymbol{Z}}_{t_{j-1}} \right) - \sum_{i=1}^{p} R_{i,t_j} \gamma_i. \tag{S1.14}$$

We proceed to prove theorem by induction on j. When $j = 1$, (S1.14) becomes

$$\begin{aligned}\hat{\boldsymbol{Z}}_{t_1} - \boldsymbol{Z}_{t_1}^{(0)} &= B^{-1}\sum_{i=1}^{p}\hat{R}_{i,t_{i1}}\hat{\gamma}_i I_{i1} - \sum_{i=1}^{p} R_{i,t_1}\gamma_i \\ &= B^{-1}\sum_{i=1}^{p}\left(\hat{R}_{i,t_{i1}}\hat{\gamma}_i - R_{i,t_j}\gamma_i\right) I_{i1} + (B^{-1} - I)\sum_{i=1}^{p} R_{i,t_1}\gamma_i I_{i1} \\ &\quad - \sum_{i=1}^{p} R_{i,t_1}\gamma_i(1 - I_{i1}).\end{aligned}$$

From (S1.12), (S1.13), and the definition of $\hat{R}_{i,t_{i1}}$, we conclude that

$$\|\hat{\boldsymbol{Z}}_{t_1} - \boldsymbol{Z}_{t_1}^{(0)}\| = O_p(n^{-1/4}) + O(\delta_n) + O(m_1),$$

where we use the fact that

$$B^{-1} - I = B^{-1}\left(I - \sum_{i=1}^{p}\hat{\gamma}_i\hat{\gamma}_i^{\top} I_{ij} - \delta_n\hat{A}^{-1}\right).$$

Suppose that (3.8) is true for $j = k - 1$. In other words,

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

holds for $j = k - 1$. Consider $j = k$ now. Similarly, we conclude from (S1.14) that

$$\begin{aligned}\|\hat{\boldsymbol{Z}}_{t_j} - \boldsymbol{Z}_{t_j}^{(0)}\| &= O_p(n^{-1/4}) + O(\delta_n) + O(m_j) + O_p(\delta_n m_{j-1}) \\ &= O_p(n^{-1/4}) + O(\delta_n) + O(m_j).\end{aligned}$$

The proof is complete. □

S1.4 Proof of Theorem 4

Let S_{true} be the multiple pre-averaging estimation based on true but unobservable log prices. Namely

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

where $\bar{\mathbf{Y}}_{t_j}^{n,(0)} = \frac{1}{k_n}\left(\sum_{\ell=k_n/2}^{k_n-1}\boldsymbol{Y}_{t_{j+\ell}} - \sum_{\ell=0}^{k_n/2}\boldsymbol{Y}_{t_{j+\ell}}\right)$. In the following, we will show that S_1 of equation (3.9) converges in probability to S_{true}, and hence, S_1 has the same limiting distribution as S_{true}.

By using the alternative version of $\bar{\mathbf{Y}}_{t_j}^{n}$ shown in Jacod et al. (2009), given $g(x) = \min\{x, 1-x\}$, we have

$$\begin{aligned}
\bar{\mathbf{Y}}_{t_j}^{n} &= \sum_{\ell=1}^{k_n-1} g(\ell/k_n)(\hat{\boldsymbol{Y}}_{t_{j+\ell}} - \hat{\boldsymbol{Y}}_{t_{j+\ell-1}}) \\
&= \sum_{\ell=1}^{k_n-1} g(\ell/k_n)(\hat{\boldsymbol{X}}_{t_{j+\ell}} - \hat{\boldsymbol{X}}_{t_{j+\ell-1}} + \hat{\boldsymbol{\varepsilon}}_{t_{j+\ell}} - \hat{\boldsymbol{\varepsilon}}_{t_{j+\ell-1}}) \\
&= \sum_{\ell=1}^{k_n-1} g(\ell/k_n)(\boldsymbol{X}_{t_{j+\ell}} - \boldsymbol{X}_{t_{j+\ell-1}}) + k_n^{-1}\sum_{\ell=k_n/2+1}^{k_n-1}\boldsymbol{\varepsilon}_{t_{j+\ell}} - k_n^{-1}\sum_{\ell=1}^{k_n/2}\boldsymbol{\varepsilon}_{t_{j+\ell}} \\
&\quad + \sum_{\ell=1}^{k_n-1} g(\ell/k_n)(\hat{\boldsymbol{X}}_{t_{j+\ell}} - \boldsymbol{X}_{t_{j+\ell}} - (\hat{\boldsymbol{X}}_{t_{j+\ell-1}} - \boldsymbol{X}_{t_{j+\ell-1}})) \\
&\quad + k_n^{-1}\sum_{\ell=k_n/2+1}^{k_n-1}(\hat{\boldsymbol{\varepsilon}}_{t_{j+\ell}} - \boldsymbol{\varepsilon}_{t_{j+\ell}}) - k_n^{-1}\sum_{\ell=1}^{k_n/2}(\hat{\boldsymbol{\varepsilon}}_{t_{j+\ell}} - \boldsymbol{\varepsilon}_{t_{j+\ell}}) \\
&= \bar{\mathbf{Y}}_{t_j}^{n,(0)} + L_1 + L_2, \qquad \text{(S1.15)}
\end{aligned}$$

where

$$L_1 = \sum_{\ell=1}^{k_n-1} g(\ell/k_n)(\hat{\boldsymbol{X}}_{t_{j+\ell}} - \boldsymbol{X}_{t_{j+\ell}} - (\hat{\boldsymbol{X}}_{t_{j+\ell-1}} - \boldsymbol{X}_{t_{j+\ell-1}})),$$
$$L_2 = k_n^{-1} \sum_{\ell=k_n/2+1}^{k_n-1} (\hat{\boldsymbol{\varepsilon}}_{t_{j+\ell}} - \boldsymbol{\varepsilon}_{t_{j+\ell}}) - k_n^{-1} \sum_{\ell=1}^{k_n/2} (\hat{\boldsymbol{\varepsilon}}_{t_{j+\ell}} - \boldsymbol{\varepsilon}_{t_{j+\ell}}).$$

Now we will verify the orders of $\bar{\mathbf{Y}}_{t_j}^{n,(0)}$, L_1, and L_2.

Since $S_{\text{true}} - \Sigma = O_p(n^{-1/4})$, we obtain the order $\alpha = 1/4$ in equation (3.8) by Theorem 3. Then, for all $j = 1, \ldots, n$,

$$\hat{\mathbf{R}}_{t_j} - \mathbf{R}_{t_j} = (\hat{\Gamma} - \Gamma)^\top (\hat{\boldsymbol{Z}}_{t_j} - \boldsymbol{Z}_{t_j}) + \Gamma^\top (\hat{\boldsymbol{Z}}_{t_j} - \boldsymbol{Z}_{t_j}) + (\hat{\Gamma} - \Gamma)^\top \boldsymbol{Z}_{t_j} = o_p(n^{-1/8+\eta/2}),$$

whereas $\hat{\mathbf{R}}_{t_j} - \mathbf{R}_{t_j} = \hat{\boldsymbol{X}}_{t_j} - \boldsymbol{X}_{t_j} + \hat{\boldsymbol{\varepsilon}}_{t_j} - \boldsymbol{\varepsilon}_{t_j}$. By using assumption (i), it implies that

$$\hat{\boldsymbol{X}}_{t_j} - \boldsymbol{X}_{t_j} = o_p(n^{-1/8+\eta/2}) \text{ and } \hat{\boldsymbol{\varepsilon}}_{t_j} - \boldsymbol{\varepsilon}_{t_j} = o_p(n^{-1/8+\eta/2}). \tag{S1.16}$$

Therefore, by (S1.15), (S1.16), the orders of $\bar{\mathbf{Y}}_{t_j}^{n,(0)}$ and L_1 are

$$\begin{aligned} \bar{\mathbf{Y}}_{t_j}^{n,(0)} &= \sum_{\ell=1}^{k_n-1} g(\ell/k_n)(\boldsymbol{X}_{t_{j+\ell}} - \boldsymbol{X}_{t_{j+\ell-1}}) + k_n^{-1} \sum_{\ell=k_n/2+1}^{k_n-1} \boldsymbol{\varepsilon}_{t_{j+\ell}} - k_n^{-1} \sum_{\ell=1}^{k_n/2} \boldsymbol{\varepsilon}_{t_{j+\ell}} \\ &= O(\sqrt{k_n}) O_p(n^{-1/2}) + O(k_n^{-1}) O_p(1) = O_p(n^{-1/4}), \\ L_1 &= \sum_{\ell=1}^{k_n-1} g(\ell/k_n)(\hat{\boldsymbol{X}}_{t_{j+\ell}} - \boldsymbol{X}_{t_{j+\ell}} - (\hat{\boldsymbol{X}}_{t_{j+\ell-1}} - \boldsymbol{X}_{t_{j+\ell-1}})), \\ &= O(\sqrt{k_n}) O_p(n^{-1/2}) o_p(n^{-1/8+\eta/2}) = o_p(n^{-3/8+\eta/2}), \end{aligned} \tag{S1.17}$$

and, by assumption (ii), the order of L_2 is

$$
\begin{aligned}
L_2 &= k_n^{-1} \sum_{\ell=k_n/2+1}^{k_n-1} (\hat{\boldsymbol{\varepsilon}}_{t_{j+\ell}} - \boldsymbol{\varepsilon}_{t_{j+\ell}}) - k_n^{-1} \sum_{\ell=1}^{k_n/2} (\hat{\boldsymbol{\varepsilon}}_{t_{j+\ell}} - \boldsymbol{\varepsilon}_{t_{j+\ell}}), \\
&= O(k_n^{-1}) o_p(n^{-1/8+\eta/2}) = o_p(n^{-5/8+\eta/2}), \qquad \text{(S1.18)}
\end{aligned}
$$

where we set $k_n = \lfloor \theta\sqrt{n} \rfloor = O(n^{1/2})$.

Next, for the second term of (3.9), we verify in the similar way.

$$
\begin{aligned}
\hat{\mathbf{Y}}_{t_j} - \hat{\mathbf{Y}}_{t_{j-1}} &= \mathbf{Y}_{t_j} - \mathbf{Y}_{t_{j-1}} + (\hat{\boldsymbol{X}}_{t_j} - \boldsymbol{X}_{t_j} - (\hat{\boldsymbol{X}}_{t_{j-1}} - \boldsymbol{X}_{t_{j-1}})) \\
&\quad + (\hat{\boldsymbol{\varepsilon}}_{t_j} - \boldsymbol{\varepsilon}_{t_j} - (\hat{\boldsymbol{\varepsilon}}_{t_{j-1}} - \boldsymbol{\varepsilon}_{t_{j-1}})) \\
&= \mathbf{R}_{t_j} + L_3 + L_4, \qquad \text{(S1.19)}
\end{aligned}
$$

where the orders of $\mathbf{R}_{t_j}$, L_3, and L_4 are $O_p(1)$, $o_p(n^{-5/8+\eta/2})$ and $o_p(n^{-1/8+\eta/2})$, respectively.

Finally, combining (S1.15), (S1.17), (S1.18) and (S1.19), we have

$$
\begin{aligned}
\|n^{1/4}(S_1 - S_{true})\| &= n^{1/4}\left\| \frac{n}{n-k_n+2}\frac{12}{k_n}\sum_{j=0}^{n-k_n+1}\left[\bar{\mathbf{Y}}_{t_j}^{n,(0)}(L_1+L_2)+(L_1+L_2)^2\right]\right. \\
&\quad \left. -\frac{12}{2n\theta^2}\sum_{j=1}^{n}\left[\mathbf{R}_{t_j}(L_3+L_4)+(L_3+L_4)^2\right]\right\| \\
&\leq n^{1/4}\frac{n}{n-k_n+2}\frac{12}{k_n}\left\|\sum_{j=0}^{n-k_n+1}\bar{\mathbf{Y}}_{t_j}^{n,(0)}(L_1+L_2)\right\| \\
&\quad +n^{1/4}\frac{n}{n-k_n+2}\frac{12}{k_n}\left\|\sum_{j=0}^{n-k_n+1}(L_1+L_2)^2\right\| \\
&\quad +n^{1/4}\frac{12}{2n\theta^2}\left\|\sum_{j=1}^{n}\left[\mathbf{R}_{t_j}(L_3+L_4)+(L_3+L_4)^2\right]\right\| \\
&\leq O(n^{1/2}k_n^{-2})O(n)O_p(n^{-1/4})o_p(n^{-3/8+\eta/2}) \\
&\quad +O(n^{1/2}k_n^{-2})O(n)o_p(n^{-3/4+\eta}) \\
&\quad +O(n^{1/2}n^{-2})O(n)o_p(n^{-1/8+\eta/2}) \\
&= o_p(1).
\end{aligned}
$$

This validates the theorem. □

S2 Relevant Simulations

We propose an eigenvalue correction method (Section 3.1) and a synchronization technique (Section 3.2). In what follows, we confirm our theoretical results by using the limiting distributional property of the proposed eigenvalue correction method, and compare this method with alternatives, such as replacing negative eigenvalues with small positive values (McNeil et al., 2005; Schaeffer, 2014) or with zeros (Rebonato and Jäckel, 1999) (Section S2.1). We also consider the finite sample performance of the high frequency filtration technique relative to the alternative previous tick and refresh time synchronization techniques (Section S2.2). Finally, simulation results for $p = 10$ and $p = 15$ are discussed in Section S2.3.

S2.1 Property of eigenvalue correction

We first assess the limiting distributional property of the estimator using the proposed eigenvalue correction method. As shown in Section 3.1, the proposed eigenvalue correction realized covariance estimator S has the same limiting distribution as its preliminary estimator S_1, which is efficient but may not be semi-positive definite. In this section, we use the same model as in Section 4.1 to generate synchronous yet noisy log prices. To obtain negative-definite estimated covariance matrices, we only consider the Negative scenario, for which the parameters are listed in Table S4.

For the preliminary estimator S_1, we use the multiple pre-averaging estimator (MPA) as in Christensen et al. (2010). Based on 1 000 replications, 23% of the preliminary covariance estimators are not semi-positive definite. Given the preliminary MPA estimator, we evaluate the eigenvalue correction estimator S according to (3.2). The chi-square goodness-of-fit statistic is used to test the null hypothesis that all elements of $n^{1/4}(S_1 - \Sigma)$ and $n^{1/4}(S - \Sigma)$ have the same distribution. The p-values are obtained as

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0.99 & 0.99 \\ 1 & 1 & 1 & 0.97 & 1 \\ 1 & 0.99 & 0.96 & 0.46 & 0.99 \\ 1 & 0.99 & 1 & 0.99 & 1 \end{bmatrix}$$

providing strong evidence to accept this hypothesis.

We also consider alternative eigenvalue correction approaches in which each negative eigenvalue is replaced by a zero Rebonato and Jäckel (1999) or by a small positive number McNeil et al. (2005). From (3.1), the spectral decompositions of the preliminary estimator S_1 and the true covariance matrix Σ are given by

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

where $\hat{\lambda}_i$ and λ_i 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. Note that $\hat{\lambda}_i$, $i = 1, \ldots, p$, may be negative numbers, whereas λ_i, $i = 1, \ldots, p$, are all positive numbers.

The corrected covariance matrices obtained by replacing each negative eigenvalue with its absolute value (the proposed eigenvalue correction), by a zero (Rebonato and Jäckel, 1999), and by a small positive number (McNeil et al., 2005) are denoted respectively by

$$S = U|\hat{\Lambda}|U^*, \qquad Z = U\hat{\Lambda}^+U^*, \qquad P = U\hat{\Lambda}^{(c)}U^*,$$

where $|\hat{\Lambda}| = \text{diag}(|\hat{\lambda}_1|, \ldots, |\hat{\lambda}_p|)$, $\hat{\Lambda}^+ = \text{diag}(\max\{\hat{\lambda}_1, 0\}, \ldots, \max\{\hat{\lambda}_p, 0\})$, and $\hat{\Lambda}^{(c)} = \text{diag}(\max\{\hat{\lambda}_1, c\}, \ldots, \max\{\hat{\lambda}_p, c\})$. In this case, we specify a data-driven constant c given by the smallest positive eigenvalue divided by 2; that is, $c = \min\{\{\max\{\hat{\lambda}_1, 0\}, \ldots, \max\{\hat{\lambda}_p, 0\}\} \setminus \{0\}\}/2$.

We evaluate the performance of these corrected covariance matrices by computing the mean square error (MSE) of the negative eigenvalues, defined as

$$MSE_i^{(r)} = \frac{\sum_{s=1}^m \left[(\hat{\lambda}_i^{(r)} - \lambda_i)^2\right] \mathbf{1}\{\hat{\lambda}_i^{(s)} < 0\}}{\sum_{s=1}^m \mathbf{1}\{\hat{\lambda}_i^{(s)} < 0\}}, \quad i = 1, \ldots, p,$$

where $\hat{\lambda}_i^{(s)}$ denotes the i-th estimated eigenvalue of S_1 for the s-th replication, and $\hat{\lambda}_i^{(r)}$, $r = S,\ Z,\ P$, denotes the i-th corrected eigenvalue of $S,\ Z$, and P, respectively.

For the Negative scenario with $p = 5$, the estimator S_1 is not positive-semidefinite in 230 out of the 1 000 replicates. The negative eigenvalues mostly occur for the fifth eigenvalue. The MSEs for S, Z, and P and the relative efficiencies with respect to S are shown in Table S1. The proposed approach is observed to improve the MSEs of the smallest eigenvalues by 43% and 67% compared to the methods of Rebonato and Jäckel (1999) and McNeil et al. (2005), respectively. We note that there may exist an alternative data-driven method to determine c that would improve the performance of P. However, empirical results indicate that such a method is not easy to obtain.

A potential concern regarding the proposed eigenvalue correction approach is that a large negative eigenvalue may be replaced with a large positive one. In reality, however, the population eigenvalues (λ_i) are all positive and there are no large sample negative eigenvalues. To verify this assertion, we consider an experiment with $p = 10$ in which the parameter settings are doubled the values of all parameters of the Negative scenario with $p = 5$. In particular, we repeat the procedures outlined above with $p = 10$. The results are shown in Table S2.

On the basis of these results, the proposed eigenvalue correction method still performs best for the smallest eigenvalue ($i = 10$), and for the third smallest eigenvalue ($i = 8$). However, it is outperformed by the method of McNeil et al. (2005) for the second smallest eigenvalue ($i = 9$). This phenomenon may be due to a choice of the constant c which is suitable for a particular eigenvalue (e.g. $i = 9$ in this experiment), but not for all eigenvalues. Conversely, the proposed eigenvalue correction method is suitable for all of the estimated negative eigenvalues.

Table S1: The mean square errors ($\times 10^{-7}$) of the negative eigenvalue and the corresponding relative efficiencies in the case $p = 5$.

	S	Z	P
$MSE_5^{(r)}$	8.49	12.2	14.1
$\frac{MSE_5^{(r)}}{MSE_5^{(S)}}$	1	1.43	1.67

Table S2: The mean square errors ($\times 10^{-7}$) of the negative eigenvalues and the corresponding relative efficiencies in the case $p = 10$.

	$i = 8$			$i = 9$			$i = 10$		
	S	Z	P	S	Z	P	S	Z	P
$MSE_i^{(r)}$	12.4	12.8	12.8	12.1	12.2	11.8	6.43	8.47	6.67
$\frac{MSE_i^{(r)}}{MSE_i^{(S)}}$	1	1.024	1.024	1	1.004	0.973	1	1.317	1.037

S2.2 Synchronization

In this section, we compare three synchronization techniques: previous tick (PT), refresh time (RT), and the proposed high frequency filtering (HFF). Our comparison of these techniques is performed using the Electronic and

Ex-HF experiments described previously.

For the PT technique, we interpolate the asynchronous series with the average sample size $n = \sum_{i=1}^{p} n_i/p$, where n_i is the sample size of i-th dimension. We compute the MSE of the synchronized log prices with respect to the true values based on 1000 replications. It should be noted, however, that the primary goal in our study is not to look for the "missing" returns. Rather, it is to filter out the synchronous processes at high sampling frequency without destroying the original cross-dependence in the raw data.

Table S3 contains the average sample sizes of the obtained synchronous data, and the MSEs across the 1000 replications, for the three considered techniques. The results indicate that, without exception, the HFF technique provides the largest sample sizes. In the Electronic experiment, the interpolated sample size of HFF is 10 times as large as that of RT, and 3.5 times as large as that of PT, with a slightly bigger MSE. In the Ex-HF experiment, the interpolated sample size of HFF is approximately 6 times as large as that of RT, and 2.5 times as large as that of PT, with similar MSEs. Thus, while the alternative techniques deliver slightly more accurate "missing" values, they are tracing not only the efficient returns, but a combination of the efficient returns and microstructure noise, the latter of which has the potential to introduce bias in the estimation as illustrated in

Tables 1 and S5.

Table S3: Synchronization comparison.

Electronic	average sample size	MSE
Previous ticks	3900.00	1.93×10^{-6}
Refresh time	1193.29	1.56×10^{-6}
High frequency filtration	13323.30	2.08×10^{-6}
Ex-HF case	average sample size	mean
Previous ticks	7800.00	1.62×10^{-6}
Refresh time	3183.67	1.21×10^{-6}
High frequency filtration	18979.40	1.36×10^{-6}

S2.3 Simulation results on high dimensional cases

In this section, we consider the relative errors and maximum norms of the efficient multiple pre-averaging (EMP) estimator compared with multiple kernel (MK) estimator and multiple pre-averaging (MPA) estimator for the cases $p = 10$ and $p = 15$. The results are shown in Tables S5 and S6. Without exception, the EMP estimator outperforms MK and MPA in terms of the relative error for each eigenvalue. In particular, the EMP estimator delivers relative enhancements of between 75% and 279% compared to MPA, and of between 64% and 278% compared to MK. In terms of the maximum norm, the EMP estimator outperforms MPA and MK by 97% and 133%, respectively.

Figures S1 and S2 depict the relative errors and the normalized mean absolute errors for each estimated eigenvalue of MPA, MPA-E, MPA-H, and EMP. Comparing MPA-E with MPA, we observe that the proposed eigenvalue correction contributes mostly to the negative eigenvalues. Comparing MPA-H with EMP, we note that the HFF technique will interpolate too many biased observations if the preliminary covariance matrix is not positive semidefinite.

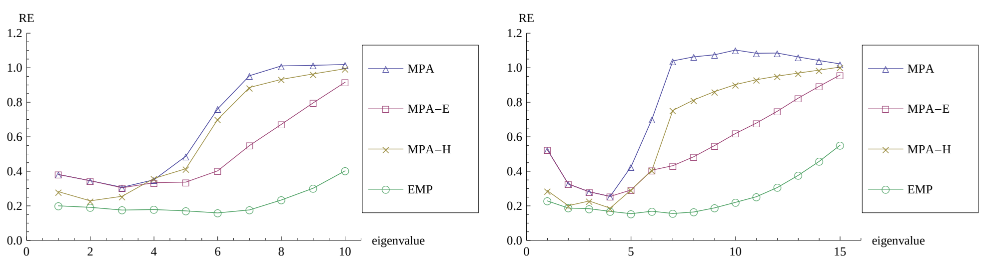

Figure S1: The relative errors (REs) of each eigenvalue based on MPA, MPA-E, MPA-H, and EMP for $p = 10$ (upper panel) and $p = 15$ (lower panel).

The computational times of MPA, MK and EMP for one replication are 178, 14 and 62 seconds, respectively, using the Mathematica 8 software on an Intel(R) Xeon(R) CPU E7-4860@2.27GHz, and a total RAM of 252 GB.

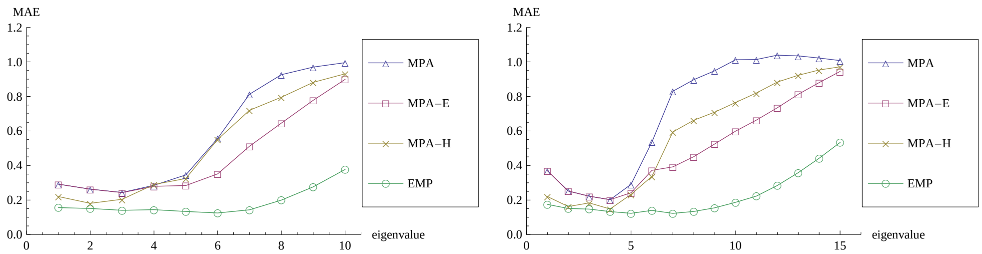

Figure S2: The mean absolute errors (MAEs) of each eigenvalue based on MPA, MPA-E, MPA-H, and EMP for $p = 10$ (upper panel) and $p = 15$ (lower panel).

Bibliography

Bai, Z. D., B. Q. Miao, and C. R. Rao (1991). Estimation of directions of arrival of signals: Asymptotic results. advances in spectrum analysis and array processing. pp. 327–347. Prentice Hall's West Nyack, New York,.

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.

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.

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.

Table S4: The parameter setting of the long-term mean of the volatility (μ_i), the variance of the microstructure noise (η_i) and the intensities (ψ).

	Finance						Noisy				
i	1	2	3	4	5	i	1	2	3	4	5
$\mu_i(\times 10^{-4})$	1.6	3.2	1.6	4.8	1.6	$\mu_i(\times 10^{-4})$	1.6	3.2	1.6	2.4	1.6
$\eta_i(\times 10^{-7})$	1	0.6	1	4	0.8	$\eta_i(\times 10^{-7})$	4	4	4	4	4
ψ_i	10	6	13	8	8	ψ_i	3	5	8	10	12
	Electronics						Ex-Asy				
i	1	2	3	4	5	i	1	2	3	4	5
$\mu_i(\times 10^{-4})$	4.8	3.2	1.6	3.2	1.6	$\mu_i(\times 10^{-4})$	4.8	3.2	4.8	3.2	1.6
$\eta_i(\times 10^{-7})$	4	4	0.4	1	0.4	$\eta_i(\times 10^{-7})$	4	1	0.8	0.6	0.4
ψ_i	3	5	8	10	12	ψ_i	3	6	10	20	60
	Food						Ex-HF				
i	1	2	3	4	5	i	1	2	3	4	5
$\mu_i(\times 10^{-4})$	3.2	4.8	2.8	1.6	1.6	$\mu_i(\times 10^{-4})$	1.6	3.2	1.6	4.8	1.6
$\eta_i(\times 10^{-7})$	2.5	4	3	0.8	0.8	$\eta_i(\times 10^{-7})$	1	4	0.6	0.8	4
ψ_i	10	6	8	12	12	ψ_i	3	3	5	5	5
							Negative				
						i	1	2	3	4	5
						$\mu_i(\times 10^{-4})$	0.4	1.6	0.16	0.04	0.16
						$\eta_i(\times 10^{-7})$	4	4	4	4	4
						ψ_i	3	5	3	5	3

Table S5: A comparison of EMP with MK and MPA in terms of relative errors (REs) of eigenvalues and maximum norms (MNs), in the case $p = 10$.

	Finance & Electronics						
	EMP	$\frac{\text{MK}}{\text{EMP}}$	$\frac{\text{MPA}}{\text{EMP}}$		EMP	$\frac{\text{MK}}{\text{EMP}}$	$\frac{\text{MPA}}{\text{EMP}}$
RE_1	0.200	2.674	1.906	RE_6	0.159	3.059	3.789
RE_2	0.191	2.060	1.806	RE_7	0.176	3.782	3.405
RE_3	0.176	1.857	1.750	RE_8	0.234	3.455	3.318
RE_4	0.179	2.029	1.966	RE_9	0.302	3.008	3.353
RE_5	0.171	2.344	2.853	RE_{10}	0.403	2.422	2.529
MN	$1.48E-4$	2.399	1.971				

Table S6: A comparison of EMP with MK and in terms of the relative errors (REs) of eigenvalues and maximum norms (MNs), in the case $p = 15$.

	Finance & Electronics & Food										
	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	0.228	2.882	2.296	RE_6	0.169	2.680	2.154	RE_{11}	0.253	2.756	2.283
RE_2	0.187	1.987	1.745	RE_7	0.156	3.257	3.677	RE_{12}	0.306	2.053	2.541
RE_3	0.183	1.638	1.534	RE_8	0.164	3.284	3.466	RE_{13}	0.377	2.669	2.821
RE_4	0.168	2.114	1.515	RE_9	0.188	3.067	2.707	RE_{14}	0.457	2.171	2.276
RE_5	0.155	2.864	2.761	RE_{10}	0.220	2.756	3.003	RE_{15}	0.551	1.846	1.855
MN	$1.86E-4$	2.338	2.063								