Efficient and Semi-Positive Definite Pre-Averaging Realized Covariance Estimator

Y Chen, LC Lin, G Pan and V Spokoiny

National University of Singapore, Singapore
National Sun Yat-Sen University, Taiwan
Nanyang Technological University, Singapore
Weierstraß Institute for Applied Analysis and Stochastic, Germany

Covariance

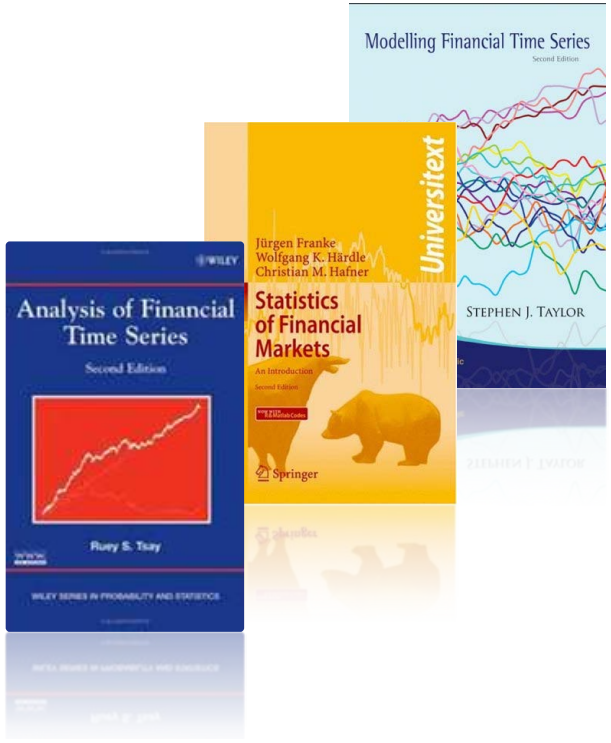
- A measure of uncertainty about returns;
- An input parameter in many financial activities such as risk management, derivative pricing, hedging and portfolio selection.

Remarks:

- Neither covariance nor its elements are directly observable in markets,
- Covariance is often estimated as a latent variable based on the historical returns.

Covariance models

- Multivariate ARCH/GARCH
- Multivariate stochastic volatility models



Ultra-high frequency (UHF) data

An increasing availability of UHF data in financial markets.

- Transactions or ticks are recorded at a high sampling frequency such as minutes, seconds or even milliseconds.
- Data contain plenty of information and can be effectively used to highlight some essential features of financial variables.

Estimate covariance from the UHF data!

Univariate case

Realized variance: sum of the squared UHF returns.

- It is asymptotically consistent, see Barndorff-Nielsen and Shephard (2002).
- It displays good performance.
  - Variance modeling and prediction, see French, Schwert and Stambaugh (1987); Andersen and Bollerslev (1998); Andersen, Bollerslev, Diebold and Labys (2001).
  - Portfolio optimization, see e.g. Fan, Li and Yu (2012).

For a systematic review, see McAleer and Medeiros (2008).

Microstructure noises

Microstructure noises generates a substantial bias in the estimation, see e.g. Andersen, Bollerslev, Diebold and Ebens (2001); Barndorff-Nielsen and Shephard (2002a); Bandi and Russell (2005a).

- Optimal sampling frequency, see Bandi and Russell (2005b).
- Multi-scaling method, see Zhang, Mykland and Aït-Sahalia (2005).
- Autocorrelations correction, see Zhou (1996); Hansen and Lunde (2006); Barndorff-Nielsen, Hansen, Lunde and Shephard (2008).
- Pre-averaging approach, see Podolskij and Vetter (2006), Jacod, Li, Mykland, Podolskij and Vetter (2009).

Realized covariance

RCoV estimators

- Multi-scaled estimator: convergence rate at $\mathcal{O}_p\left(n^{-1/6}\right)$, see Wang and Zou (2010); Zhang (2011).
- Multivariate realized kernel estimator: convergence rate at $\mathcal{O}_p\left(n^{-1/5}\right)$, see Barndorff-Nielsen, Hansen, Lunde and Shephard (2011).
- Multivariate pre-averaging estimator: convergence rate at $\mathcal{O}_p\left(n^{-1/4}\right)$, see Christensen, Kinnebrock and Podolskij (2010).

Limitation

Though these estimators are robust to the presence of microstructure noise, they cannot ensure semi-positive definition and the optimal efficiency simultaneously, after employing the bias-correction approaches.

The pre-averaging covariance estimator reduces to a sub-optimal rate at $\mathcal{O}_p\left(n^{-1/5}\right)$ to ensure semi-positive definition.

Asynchrony

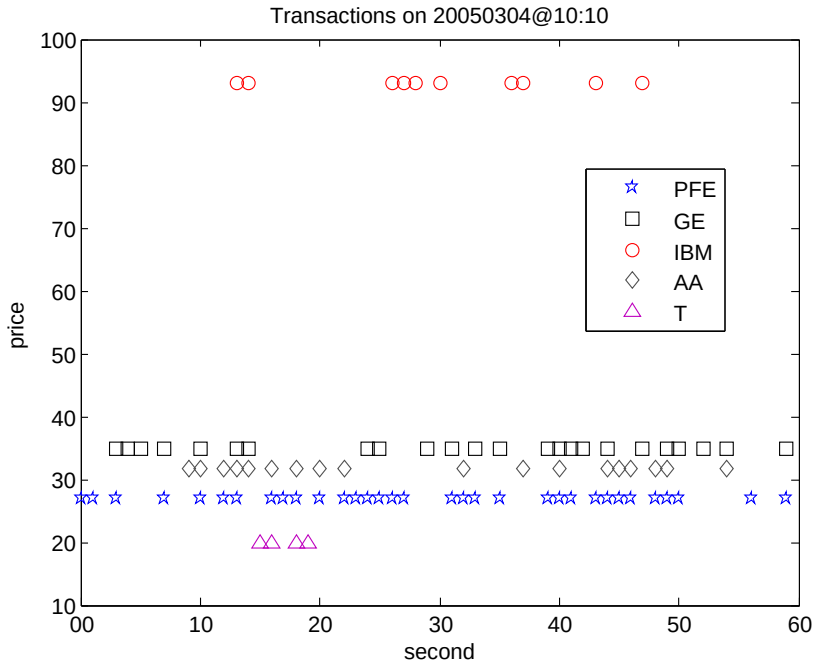


Figure 1: Transaction prices of stocks PFE, GE, IBM, AA and T on Friday, 4th March 2005@10:10:00 – 10:11:00. Data source: TAQ database.

Synchronizing techniques

- The previous tick, see e.g. Wasserfallen and Zimmermann (1985); Dacorogna, Gençay, Müller, Olsen and Pictet (2001). A spurious jump may appear many times. Distort dependence structure of raw data.
- The refresh time (RF) technique, see Hayashi and Yoshida (2005). Discard of much information that could be useful. It may yield high discretization error in the covariance estimation.

Asynchronous trading introduces negative serial correlations, even though the financial instruments are independent, see Lo and MacKinlay (1990); Goodhart (1991); Goodhart and Figliuoli (1991); Campbell, Lo and MacKinlay (1996); Campbell, Lo, MacKinlay and Whitelaw (1998).

EP Realized Covariance Estimator

Develop an **E**fficient and semi-**P**ositive definite pre-averaging realized covariance estimator.

- Asynchrony: an innovative high frequency filtering (HFF) technique
  - Data-driven synchronizing technique, learning from the dependence structure and reflecting the empirical features of raw data. ✓
- Semi-positive definiteness: a new correction approach.
  - Reach to the optimal convergence rate at $\mathcal{O}_p\left(n^{-1/4}\right)$. ✓
  - Semi-positive definite. ✓

Outline

1. Motivation ✓
2. Methods: multivariate pre-averaging estimator, HFF synchronizing technique, and SPD correction.
3. Numerical analysis
4. Conclusion

Model

Consider p assets over a time interval $[0, 1]$. The log price, $\mathbf{X}_t$, follows a continuous time diffusion model,

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

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

$$[\mathbf{X}, \mathbf{X}]_t = \int_0^t \Sigma_u du = \int_0^t \boldsymbol{\sigma}_t^T \boldsymbol{\sigma}_t du, \quad t \in [0, 1].$$

The integrated volatility matrix is denoted as Σ.

Our goal is to estimate the integrated volatility matrix Σ.

Microstructure noise

Denote $\mathbf{X}_{t_j} = (X_{1,t_j}, \ldots, X_{p,t_j})^\top$ to be the efficient log prices of the assets at time t_j, where $t_j = j/n$, $j = 0, \ldots, n$ and n refers to be the sample size at high sampling frequency. The efficient log prices are synchronous and noise-free, but unobservable in practice.

The log prices are contaminated by microstructure noise:

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

where the microstructure noise $\boldsymbol{\epsilon}_{t_j} \sim IID(0, \Psi)$. Ψ is diagonal matrix with variation of microstructure noise η_i^2, $i = 1, \ldots, p$, on the diagonal.

Autocorrelation

The lag-1 autocorrelation of the contaminated log returns ($R_{i,t_j} = Y_{i,t_j} - Y_{i,t_{j-1}}$) is

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

where $\Sigma_{ii,u}$ denotes the (i,i)-component of Σ_u of (1).

Information set of observation

The observed log prices are irregularly spaced. We define an information set $\mathcal{F}$ as:

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

which t_{ij} represents the time points t_j only when the log price $Y_{i,t_{ij}}$ of the i-th asset is observed at time t_j.

- If $t_{ij} \in \mathcal{F}$, we have $Y_{i,t_{ij}} = Y_{i,t_j}$.
- If $t_{ij} \notin \mathcal{F}$, the respective synchronous log price Y_{i,t_j} is considered as missing value and will be filtered.

The HFF synchronizing technique

Let S_0 be covariance matrix of the noisy raw data. It is the sum of the integrated covariance matrix Σ and the microstructure noise variance Ψ. We have the spectral decomposition:

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

where A is a diagonal matrix with eigenvalues a_i of S_0 on the diagonal, Γ is the associated orthonormal eigenvectors matrix.

Assume there exists a linear filter $\mathbf{Z}_{t_j}^{(0)} = \left(Z_{1t_j}^{(0)}, \ldots, Z_{pt_j}^{(0)}\right)^\top$:

$$\mathbf{R}_{t_j} = \Gamma^\top \mathbf{Z}_{t_j}, \quad t_j = j/n, \quad j = 1, \ldots, n.$$

where Γ is the eigenvector matrix.

Filtering

Starting from time t_1 to t_n, the HFF technique iteratively synchronizes data at high sampling frequency with a sequential optimization:

$$\min_{\mathbf{Z}_{t_j}} f(\mathbf{Z}_{t_j}), \tag{1}$$

where

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

$\hat{\Gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)$ and $\hat{A} = \text{diag}\{\hat{a}_1, \ldots, \hat{a}_p\}$ respectively are the eigenvectors and eigenvalues of a preliminary estimator of S_0.

Filtering

- The first part minimizes the projection errors, which is not unique due to the missing values.
- The second part reflects the existence of lag-1 autocorrelation and also ensures the synchronizing technique continues, even if the concurrent returns of some assets are not observed.

Algorithm

Set $j = 1$. Let $\hat{\mathbf{Z}}_{t_0} = \mathbf{0}_p$ and we have $\hat{S}_0 = \hat{\Gamma}\hat{A}\hat{\Gamma}^\top$.

1. If $t_{ij} \notin \mathcal{F}$ for all $i = 1, \ldots, p$, set $\hat{\mathbf{Y}}_{t_j} = \hat{\mathbf{Y}}_{t_{j-1}}$ 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 $\hat{\mathbf{Z}}_{t_j}$ that minimizes the objective function (1). We have $\hat{\mathbf{Y}}_{t_j} = \hat{\Gamma}^\top \hat{\mathbf{Z}}_{t_j} + \hat{\mathbf{Y}}_{t_{j-1}}$.
4. Stop until $j = n$; otherwise renew $j = j + 1$ and return to step 1.

Pre-averaging estimator

Given the synchronized HF data $\hat{\mathbf{Y}}_{t_j}$, the pre-averaging estimator is computed and denoted as S_1:

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

where $\bar{\mathbf{Y}}^n_{t_j} = \frac{1}{k_n}\left(\sum_{\ell=k_n/2}^{k_n-1}\hat{\mathbf{Y}}_{t_{j+\ell}} - \sum_{\ell=0}^{k_n/2}\hat{\mathbf{Y}}_{t_{j+\ell}}\right)$, the last term is bias-correction.

- The estimator S_1 is an unbiased estimator of Σ with convergence rate $\mathcal{O}_p\left(n^{-1/4}\right)$.
- By taking $k_n = \lfloor \theta n^{0.6} \rfloor$, the bias-correction term can be ignored and the estimator becomes SPD, but the convergence rate will reduce to $\mathcal{O}_p\left(n^{-1/5}\right)$.

The SPD correction

Denote the spectral decompositions of S_1 and Σ by

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

where $\hat{\lambda}$'s and λ's are respectively eigenvalues of S_1 and Σ, and $\mathbf{u}_i$ and $\mathbf{v}_i$ are orthonormal eigenvectors associated with $\hat{\lambda}_i$ and λ_i for all i.

Our SPD estimator, denoted by S, of Σ, is proposed as:

$$S = U|\hat{\Lambda}|U^*, \tag{3}$$

Consistency

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

$$S_1 - \Sigma \xrightarrow{P} 0. \tag{4}$$

Define S by (3). Then S is a consistent estimator of Σ, that is,

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

Asymptotic distribution

When reinforcing the condition on Σ, we proved that S and S_1 have the same limiting distribution.

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

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

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

Define S by (3). Then

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

Simulation

Objective: investigate the performance of the proposed estimator under various scenarios.

- We first generate synchronous data to confirm the limiting distributional property of our proposed estimator.
- Next we generate different underlying processes to compare the features of our proposed estimator to several alternatives in terms of estimation accuracy.

Simulation: synchronous data

Log price $\mathbf{X}_t$ of p assets is generated as 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 $\Sigma_t = (\Sigma_{ij,t})_{1 \le i,j \le p}$.

The diagonal elements of Σ_t follow a 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 $\mathbf{B}_t$.

Simulation: synchronous data

The off-diagonal elements follow:

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

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 standard one dimensional Brownian motion independent of $\mathbf{B}_t$ and W_{it}.

Simulation: synchronous data

We generate the synchronous but noisy log price as follows:

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

Parameter setting:

- $p = 5$, $\eta_i = 4 \times 10^{-7}$, $i = 1, \ldots, 5$,
- $(\mu_1, \ldots, \mu_5) = (4\times10^{-5}, 1.6\times10^{-5}, 1.6\times10^{-4}, 4\times10^{-5}, 1.6\times10^{-5})$, $\theta_1 = \cdots = \theta_5 = 10$, $\omega_i = \sqrt{\mu_i \theta_i}/2$.
- Sample size $n = 23\,400$ and the replications are 1 000 times.
- The preliminary estimator S_1: Pre-averaging approach. Among the 1 000 times replications, there are 222 times of negative definite covariance estimators.

Simulation: synchronous data

The Chi-square goodness-of-fit statistic is used to test the null hypothesis that all the elements of $n^{1/4}(S_1 - \Sigma)$ and $n^{1/4}(S - \Sigma)$ have the same distribution.

The result of the p-values is

$$\begin{bmatrix} 1 & 0.99 & 1 & 1 & 0.99 \\ 0.99 & 0.72 & 1 & 0.99 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0.99 & 1 & 1 & 0.99 \\ 0.99 & 0.99 & 1 & 0.99 & 0.99 \end{bmatrix}$$

which represents a strong evidence to accept that S_1 and S have the same limiting distribution.

Simulation: asynchronous data

Data generation: We use the same model as in the synchronous experiment, but construct asynchronous and noisy observations by controlling Poisson processes with 5 intensities $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_5)^\top$. The generated processes on average have $23\,400/\psi_1$ to $23\,400/\psi_5$ observations.

The parameters and intensities are respectively estimated with the real data of NYSE.

In addition, we consider one Noisy experiment with slight signals against vast noise, an extremely asynchronous (Ex-Asy) experiment with very dissimilar sampling frequency of the five assets, and an experiment with high sampling frequency of all the five assets denoted Ex-HF.

Simulation: asynchronous data

Table 1: 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
	Electronic						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

Simulation: asynchronous data

Alternatives:

- Pre-averaging estimator with the synchronous method of Hayashi and Yoshida $\hat{\Sigma}_p + HY$
- Kernel estimator with the refresh time method $\hat{\Sigma}_k + RT$
- Two-scaled estimator with the previous tick method $\hat{\Sigma}_t + PT$.

Simulation: asynchronous data

Evaluating criteria:

- Relative errors of the five eigenvalues

$$RE_i = \frac{\sqrt{\frac{1}{m}\left[\sum_{s=1}^{m}(\hat{\lambda}_i^{(s)} - \lambda_i)^2\right]}}{\lambda_i}, \quad i = 1, \ldots, 5,$$

  where λ_i is the true eigenvalues at dimension i and $\hat{\lambda}_i^{(s)}$ is the estimated eigenvalues at dimension i in the $s-$th replication, $s = 1, \ldots, m = 1\,000$.

- Maximum norm of a matrix means the largest absolute deviation of all elements of matrix.

Simulation: asynchronous data

Table 2: Finance: The relative errors of five eigenvalues for four estimators and the mean maximum norms with the standard deviation in parentheses. The number in bold-face indicates the best accuracy in each experiment.

Finance	RE_1	RE_2	RE_3	RE_4	RE_5	max norm
$\hat{\Sigma}_p + HY$	0.223	0.195	0.241	0.233	0.418	1.129(0.623)
$\hat{\Sigma}_k + RT$	**0.214**	0.221	0.233	0.348	0.575	2.430(1.624)
$\hat{\Sigma}_t + PT$	0.486	0.449	0.394	0.483	0.719	1.101(0.429)
$\hat{\Sigma}_s + HFF$	0.218	**0.159**	**0.215**	**0.200**	**0.236**	**0.702(0.412)**

Simulation: asynchronous data

Table 3: Electronic: The relative errors of five eigenvalues for four estimators and the mean maximum norms with the standard deviation in parentheses.

Electronic	RE_1	RE_2	RE_3	RE_4	RE_5	max norm
$\hat{\Sigma}_p + HY$	0.183	0.227	0.205	0.234	0.334	1.117(0.497)
$\hat{\Sigma}_k + RT$	0.467	0.402	0.406	0.440	0.669	2.589(1.535)
$\hat{\Sigma}_t + PT$	0.222	0.274	0.408	0.602	0.909	1.231(**0.316**)
$\hat{\Sigma}_s + HFF$	**0.170**	**0.145**	**0.143**	**0.230**	**0.265**	**0.672**(0.322)

Simulation: asynchronous data

Table 4: Food: The relative errors of five eigenvalues for four estimators and the mean maximum norms with the standard deviation in parentheses.

Food	RE_1	RE_2	RE_3	RE_4	RE_5	max norm
$\hat{\Sigma}_p + HY$	0.205	0.224	0.227	0.261	0.335	1.147(0.538)
$\hat{\Sigma}_k + RT$	0.483	0.419	0.369	0.529	0.658	2.606(1.702)
$\hat{\Sigma}_t + PT$	0.202	0.250	0.271	0.521	0.803	1.026(**0.319**)
$\hat{\Sigma}_s + HFF$	**0.162**	**0.142**	**0.156**	**0.183**	**0.272**	**0.698**(0.341)

Simulation: asynchronous data

Table 5: Noisy: The relative errors of five eigenvalues for four estimators and the mean maximum norms with the standard deviation in parentheses.

Noisy	RE_1	RE_2	RE_3	RE_4	RE_5	max norm
$\hat{\Sigma}_p + HY$	0.220	0.244	0.216	**0.221**	0.369	0.866(0.409)
$\hat{\Sigma}_k + RT$	0.509	0.442	0.364	0.447	0.668	1.819(1.186)
$\hat{\Sigma}_t + PT$	0.274	0.519	1.536	1.359	1.173	1.181(0.259)
$\hat{\Sigma}_s + HFF$	**0.176**	**0.165**	**0.341**	0.253	**0.303**	**0.511(0.216)**

Simulation: asynchronous data

Table 6: Ex-Asy: The relative errors of five eigenvalues for four estimators and the mean maximum norms with the standard deviation in parentheses.

Ex-Asy	RE_1	RE_2	RE_3	RE_4	RE_5	max norm
$\hat{\Sigma}_p + HY$	0.217	0.304	0.273	0.333	0.866	1.634(0.723)
$\hat{\Sigma}_k + RT$	0.608	0.487	0.442	0.627	0.837	3.625(2.144)
$\hat{\Sigma}_t + PT$	0.247	0.552	1.937	0.516	0.593	2.613(0.572)
$\hat{\Sigma}_s + HFF$	**0.209**	**0.172**	**0.217**	**0.222**	**0.339**	**0.908(0.426)**

Simulation: asynchronous data

Table 7: Ex-HF: The relative errors of five eigenvalues for four estimators and the mean maximum norms with the standard deviation in parentheses.

Ex-HF	RE_1	RE_2	RE_3	RE_4	RE_5	max norm
$\hat{\Sigma}_p + HY$	0.183	0.188	**0.189**	0.185	0.268	1.625(0.687)
$\hat{\Sigma}_k + RT$	0.404	0.367	0.373	0.405	0.604	3.759(2.322)
$\hat{\Sigma}_t + PT$	0.239	0.249	0.684	0.877	0.967	2.596(0.541)
$\hat{\Sigma}_s + HFF$	**0.138**	**0.109**	0.256	**0.143**	**0.168**	**0.913(0.449)**

Simulation: Computational time

The computational time of $\hat{\Sigma}_p + HY$, $\hat{\Sigma}_k + RT$, $\hat{\Sigma}_t + PT$ and $\hat{\Sigma}_s + HFF$ (with 5 recursion) for one replication are 178, 14, 20, 62 seconds, respectively, by using the Mathematica 8 software based on the Intel(R) Xeon(R) CPU E7-4860@2.27GHz and total RAM to be 252 GB.

Real data analysis

We consider the TAQ data of seven assets listed on the NYSE (New York Stock Exchange): AIG, GE, IBM, JPM, MRK, PFE, and T, over the period 2005/01/02∼2005/12/31. The normal trading hours of the NYSE is 6.5 hours (23400 seconds) from 9:30 to 16:00. We remove 7 days of 2005/11/21∼2005/11/30 due to the unavailable data of asset T. There are 245 trading days left.

We construct portfolios for every transaction day, where the covariance is estimated using the high frequency data. The optimal weights are obtained as in Fan, Li and Yu (2012):

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

Real data analysis

The portfolio constructed with the proposed RCoV estimator has positive median. Moreover, the portfolio has the smallest standard deviation.

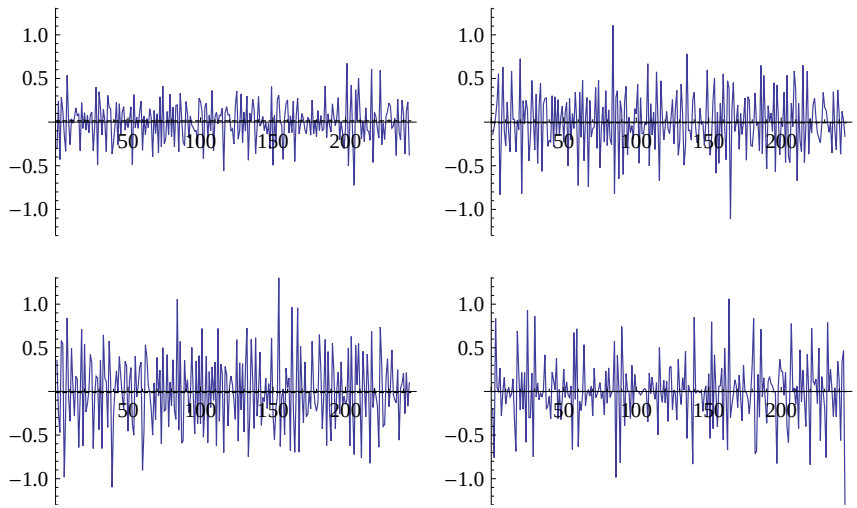


Figure 2: Time plots of the log returns of portfolio prices based on $\hat{\Sigma}_s + HFF$ (left upper panel), portfolio prices based on $\hat{\Sigma}_p + HY$ (right upper panel), portfolio prices based on $\hat{\Sigma}_k + RT$ (left lower panel), portfolio prices based on $\hat{\Sigma}_t + PT$ (right lower panel) for total 245 days.

Conclusion

We proposed an efficient and semi-positive definite pre-averaging realized covariance estimator:

- The estimator reaches the fastest convergence rate of $\mathcal{O}_p\left(n^{-1/4}\right)$. It is robust to the presence of market microstructure noise and is computed with asynchronous and noisy high frequency data.
- Asynchrony: an innovative high frequency filtering (HFF) technique that learns from the dependence structure of raw data.
- Semi-positive definiteness: a new correction approach.

Simulation study and real data analysis demonstrate superior performance compared with several alternatives.