Regularized Pairwise Estimator of Realized Covariance

Ying Chen & Vladimir Spokoiny

National University of Singapore

Department of Statistics & Applied Probability

Risk Management Institute

Weierstraß Institute for Applied Analysis and Stochastic

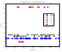
Germany

Covariance

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

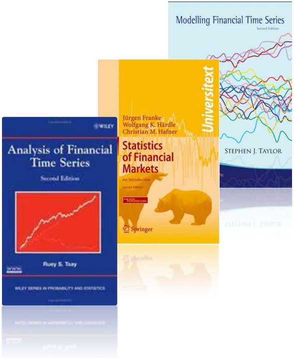
Remarks:

- 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

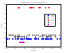


Ultra-high frequency (UHF) data

An increasing availability of UHF data in financial markets.

- Transactions or ticks are recorded at a high sampling frequency such as secondly or minutely.
- 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.

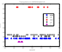
- It is asymptotically consistent, see Barndorff-Nielsen and Shephard (2002b).
- It displays a good performance.
  - Variance prediction, see French, Schwert and Stambaugh (1987); Andersen and Bollerslev (1998); Andersen, Bollerslev, Diebold and Labys (2001).
  - Portfolio optimization, see e.g. Fan, Li and Yu (2010).

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

Realized covariance

Challenges:

- **Asynchrony**: raw data are irregularly spaced and collected at different time point with different sampling frequency.
- **Microstructure noises** such as bid-ask bounce effects and price discreteness. As the sampling frequency increases, microstructure noises accumulate. It generates a substantial bias in the covariance estimation.
- **Semi-positive definiteness**: a covariance estimator should be semi-positive definite.



Asynchrony

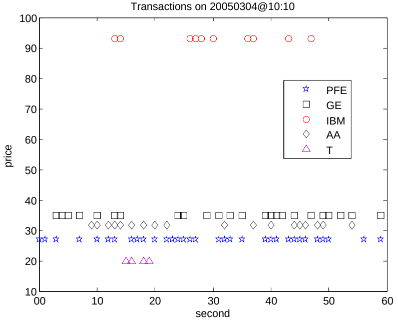
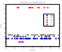


Figure 1: Transaction prices of stocks PFE, GE, IBM, AA and T on Friday, 4th March 2005@10:10:00 – 10:11:00. Data source: TAQ database.

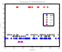
Synchronizing techniques

- The previous tick (PT) technique specifies a set of time points and takes the most recent observation for each of the time points, see e.g. Wasserfallen and Zimmermann (1985); Dacorogna, Gençay, Müller, Olsen and Pictet (2001).
- The refresh time (RF) technique picks up the time points when all the stocks were traded since last time. The last transaction of each stock is then used to construct a synchronous observation for the time point, see Hayashi and Yoshida (2005).



If some stock was traded at a low frequency …

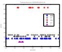
- PT: many repetitions of a particular tick.
  - A spurious jump may appear many times, which further spoils the covariance estimation.
- RF: discard of much information that could be useful. It may yield high discretization error in the covariance estimation.



Microstructure noises

Microstructure noises generates a substantial bias in the covariance 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).
- Autocorrelations correction, see Barndorff-Nielsen, Hansen, Lunde and Shephard (2008); Zhou (1996); Hansen and Lunde (2006).
- Multi-scaling method, see Zhang (2010); Zhang, Mykland and Aït-Sahalia (2005).

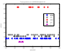


Semi-positive definiteness

- Barndorff-Nielsen et al. (2008): kernel-based estimator.
- Zhang (2010): multi-scaled estimator.
- Wang and Zou (2010): high-dimensional estimator.

Regularized estimator:

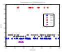
Hautsch, Kyj and Oomen (2009): blockwise kernel-based estimator, where an eigenvalue-cleaning regularization is used to guarantee the semi-positiveness.



Regularized pairwise estimator

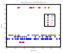
Develop a new methodology to estimate realized covariance.

- Asynchrony: high frequency filtering (HFF) technique. ✓
  - HFF is a data-driven synchronizing technique that learns from the dependence structure of raw data.
- Microstructure noises: covariance is pairwise estimated via the multi-scaling method. ✓
- Semi-positive definiteness: a regularization. ✓



Outline

1. Motivation ✓
2. Methods: HFF, multi-scaling and regularization
3. Numerical analysis
4. Conclusion



Notation

Underlying log prices $\mathbf{P}_t^* = \left(P_{1t}^*, \cdots, P_{dt}^*\right)^\top$, $t \in [0, T]$.

- The efficient log prices follow a semi-martingale process:

$$\mathbf{P}_t^* = \int_0^t \mu_s ds + \int_0^t \Theta_s dW_s$$

  where μ_t is a drift vector, Θ_t is an instantaneous co-volatility process and $\mathbf{W}_t$ is a Brownian motion.

- **Integrated covariance:** $\Sigma = \int_0^T \Theta_t \Theta_t^\top dt$.

Raw data: $\mathbf{P} = (P_{1t^{(1)}}, \cdots, P_{dt^{(d)}})$, with $t^{(j)} \in \mathcal{F}$:

$$\mathcal{F} = \left\{ t^{(j)} | P_{jt} \text{ is available at } t,\ t \in [0, T],\ j = 1, \cdots, d. \right\}$$

Synchronization: HFF technique

Let $\mathbf{X}_t^* = \mathbf{P}_t^* - \mathbf{P}_{t-1}^*$ denote the returns of the underlying synchronous series.

Suppose the covariance Σ of the synchronous returns is given:

$$\Sigma = U \Lambda U^\top$$

where Λ and $U = (U_1, \cdots, U_d)^\top$ are eigenvalue and eigenvector matrices respectively, $U^{-1} = U^\top$.

Linear transformation: project into the direction along which the underlying return series has maximum variation:

$$\mathbf{X}_t^* = U\mathbf{Z}_t, \quad \text{or} \quad \mathbf{Z}_t = U^\top \mathbf{X}_t^*.$$

Synchronization: HFF technique

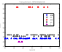
The observed log returns of the jth stock can be computed:

$$X_{jt} = \frac{P_{jt} - P_{js}}{t - s}, \quad \text{where } s \le t \text{ and } s, t \in \mathcal{F}.$$

The HFF technique is to filter out Z_t that minimizes the Euclidean distance between the filtered synchronous returns and the actual values

$$\min \sum_{j=1}^{d} \sum_{t \in \mathcal{F}} \left\{ |X_{jt} - U_j Z_t|^2 \right\}.$$

No unique solution!



Synchronization: HFF technique

Assumption: the linear filter $\mathbf{Z}_t$ is smooth,

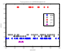
$$\widetilde{\mathbf{Z}}_t = \operatorname{argmin} \sum_{j=1}^{d} \sum_{t \in \mathcal{F}} \{|X_{jt} - U_j Z_t|^2\} + \delta \sum_{j=1}^{d} \sum_{s=1}^{T} \{Z_{js} - Z_{j,s-1}\}^2 / \lambda_j,$$

- the first part measures the Euclidean distance;
- the second part penalizes non-smoothness, measured by an instantaneous variations of Z_j standardized by its variance – the corresponding eigenvalues λ_j;
- δ controls the smoothness of the filtered series. The larger the value, the smoother the filtered series.

Synchronization: HFF technique

Remarks:

- The HFF technique filters out $\mathbf{Z}_t$ iteratively by learning from the past filtration.
- The HFF technique benefits from the usage of covariance.
- In practice, covariance is unobservable. However, an estimator based on low sampling frequency data or other covariance proxies can be used.



Removing impact of microstructure noises

Now the synchronous log prices $\mathbf{P}_t \in \mathbb{R}^d$ *are available.*

Under the presence of microstructure noises, we have:

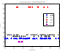
$$\mathbf{P}_t = \mathbf{P}_t^* + \varepsilon_t, \quad t = 0, \cdots, T$$

where $\mathbf{P}_t^*$ are the underlying efficient log prices and $\varepsilon_t \sim (0, \Sigma_\varepsilon)$.

The integrated covariance = the sum of the squared returns?

$$\widetilde{\Sigma} = \sum_{t=1}^{T} (P_{it} - P_{it-1})(P_{it} - P_{it-1})^\top = \Sigma + 2T\,\mathrm{E}(\varepsilon^2) + O_p(T^{1/2})$$

The bias increases with respect to the sample size T.



Removing impact of microstructure noises

Multi-scaling method: splits the entire sample to Q non-overlapping subsamples, and averages out the bias.

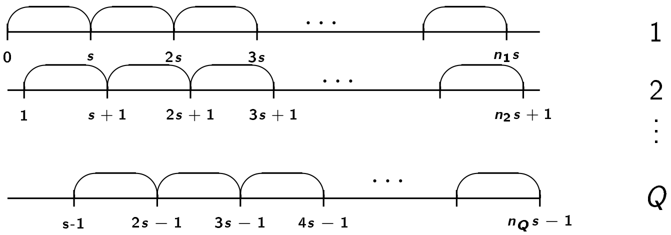


Figure 2: Multi-scaling: partition

Removing impact of microstructure noises

Let σ_{ij} denote the element of covariance Σ, we have:

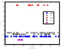
$$\tilde{\sigma}_{ij}^{(T)} = \sum_{t=1}^{T} (P_{it} - P_{it-1})(P_{jt} - P_{jt-1}) = \sigma_{ij} + 2T\,\mathrm{E}(\varepsilon^2) + O_p(T^{1/2}).$$

Analogously, we obtain other estimators based on the subsamples:

$$\tilde{\sigma}_{ij}^{(q)} = \sum_{k=q+s}^{n_q \times s+1} (P_{ik} - P_{ik-s})(P_{jk} - P_{jk-s}) = \sigma_{ij} + 2n_q\,\mathrm{E}(\varepsilon^2) + O_p(n_p^{1/2}).$$

The pairwise estimator is defined as follows:

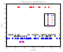
$$\tilde{\sigma}_{ij} = \frac{1}{Q}\sum_{q=1}^{Q} \tilde{\sigma}_{ij}^{(q)} - \frac{\bar{n}}{T}\tilde{\sigma}_{ij}^{(T)}, \quad \text{where } \bar{n} = \frac{1}{Q}\sum n_q.$$



Removing impact of microstructure noises

Remarks:

- The pairwise estimator is consistent and asymptotically unbiased, if the noise is IID, see Zhang (2010).
- It is empirically robust to the value of s or Q, see Zhang et al. (2005).
- However, the pairwise estimator is **not guaranteed to be semi-positive definite**.



Regularization: semi-positive definition

We are looking for a well-conditioned covariance matrix Ω that is close to the possibly not semi-positive pairwise estimator $\widetilde{\Sigma}$.

$$\min_{\Omega,\epsilon} \left\{\epsilon | \Omega \geq 0, \quad w_{ij}|\Omega_{ij} - \widetilde{\Sigma}_{ij}| \leq \epsilon, \quad 1 \leq i,j \leq p\right\}$$

$$\min_{\Omega,\epsilon} \left\{\epsilon | \Omega \geq 0, \quad \sum_{i,j=1}^{p} w_{ij}^2 \left(\Omega_{ij} - \widetilde{\Sigma}_{ij}\right)^2 \leq \epsilon, \quad 1 \leq i,j \leq p\right\}$$

$$\text{or } \min_{\Omega,\epsilon} \left\{\epsilon | \Omega \geq 0, \quad \sum_{i,j=1}^{p} w_{ij}|\Omega_{ij}\widetilde{\Sigma}_{ij}| \leq \epsilon, \quad 1 \leq i,j \leq p\right\}$$

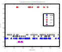
where $w_{ij} > 0$. Solving the optimization problem generates a regularized pairwise estimator.

Simulation

Objective: investigate the performance of the HFF technique.

Asynchronous data were generated based on real life UHF data – minutely transaction prices of PFE, GE, IBM, AA and T on March 4, 2005.

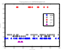
d series $\sim N_d(0, \Sigma)$, among which 1 series is re-sampled every $s > 1$ time units.



Simulation

Setup:

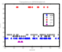
- dimensionality: $d = 2, 3, \cdots, 5$;
- sampling frequency: $s = 5, 10, 20$;
- dependence structure: Σ
  - Medium – realized covariance estimated. For example, 0.53 for $d = 2$ and a range of $[0.31, 0.53]$ for $d = 5$.
  - Low – low correlations with 0.27 for $d = 2$ and a range of $[0.16, 0.27]$ for $d = 5$.
  - High – high correlations with 0.80 for $d = 2$ and a range of $[0.59, 0.80]$ for $d = 5$.



Simulation

Average RMSE (%) of the synchronized series

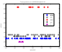
		HFF				PT
ρ	s	$d = 2$	$d = 3$	$d = 4$	$d = 5$	
low	5	**0.97**	**1.05**	**1.14**	**1.15**	1.18
low	10	**1.10**	1.18	1.25	1.23	1.16
low	20	1.23	1.24	1.21	1.18	1.13
medium	5	**0.87**	**0.94**	**1.01**	**1.01**	1.18
medium	10	**0.90**	**0.96**	**1.02**	**0.98**	1.15
medium	20	**1.01**	**1.03**	**1.04**	**1.00**	1.13
high	5	**0.71**	**0.75**	**0.78**	**0.76**	1.18
high	10	**0.65**	**0.67**	**0.67**	**0.63**	1.15
high	20	**0.69**	**0.70**	**0.69**	**0.67**	1.12



Simulation

In most cases, the HFF technique performs better than the previous tick technique.

- Dependence Σ has a substantial influence on the HFF technique. The higher dependence, the HFF technique delivers more accurate results, and vice versa.
- Dimensionality d has less effect.
- The ratio of RMSEs between the HFF technique and the previous tick technique decreases against the sampling frequency s.



Real data analysis

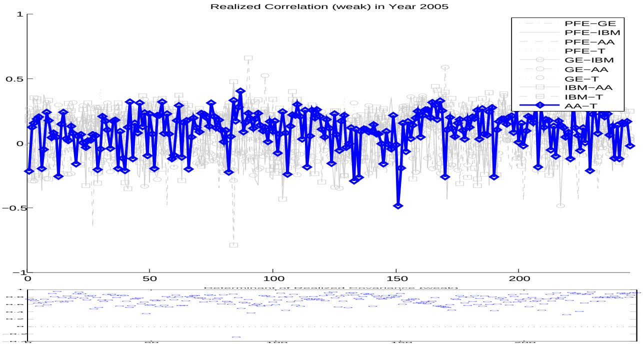


Figure 3: Realized correlation estimators for assets PFE, GE, IBM, AA and T in year 2005.

Real data analysis

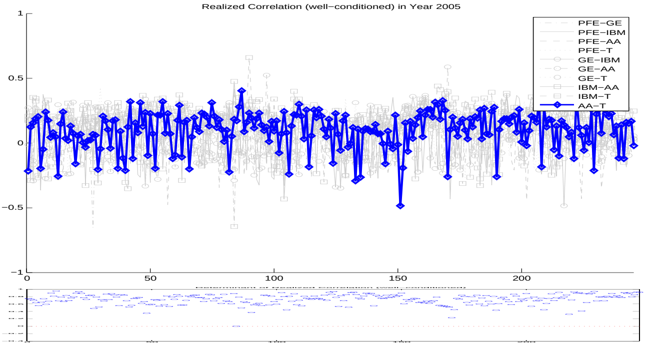


Figure 4: Realized correlation estimators for assets PFE, GE, IBM, AA and T in year 2005.

Conclusion

Develop regularized pairwise estimator – a new methodology to estimate realized covariance.

- Asynchrony: high frequency filtering (HFF) technique. ✓
  - HFF is a data-driven synchronizing technique that learns from the dependence structure of raw data.
- Microstructure noises: covariance is pairwise estimated via the multi-scaling method. ✓
- Semi-positive definiteness: a regularization. ✓