Vast Volatility Matrix Estimation for High Frequency Data

Yazhen Wang

National Science Foundation

Yale Workshop, May 14-17, 2009

Disclaimer: My opinion, not the views of NSF

Outline

High-Frequency Finance

High-Frequency Data: Intradaily observations on asset prices such as tick by tick stock price data and minute by minute exchange rate data.

Data Characteristics: High-frequency data have complex structure with microstructure noise.

One-Dim Model: Observed data: Y_{t_i}, $i = 1, \cdots, n$ and X_t = true log-price of a stock

$$\mathbf{Y_{t_i}} = \mathbf{X_{t_i}} + \epsilon_{\mathbf{t_i}}, \qquad i = 1, \cdots, n$$

ϵ_{t_i}: microstructure noise and independent of X_t.

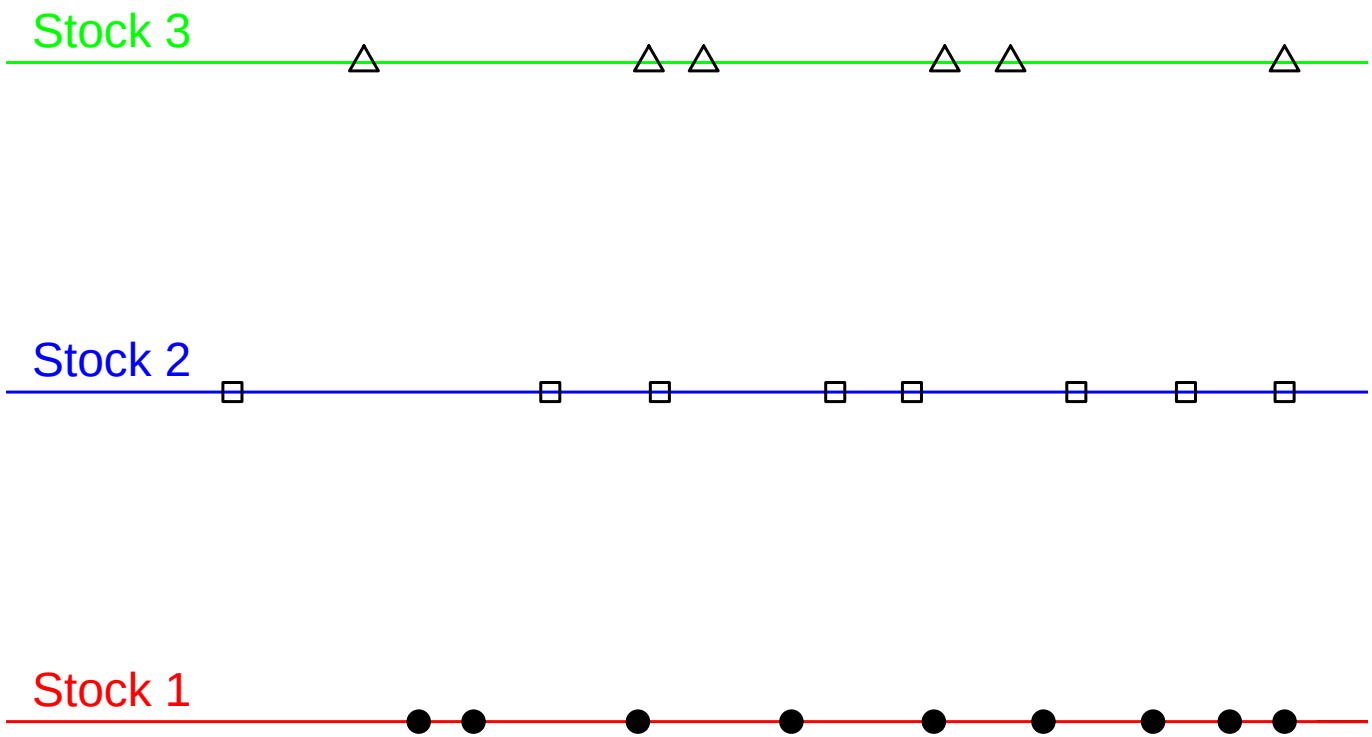
Very High Dim: Large Volatility Matrix

High Dim Model: For the i-th asset, observation times t_{ij}, $i = 1, \cdots, p$, $j = 1, \cdots, n_i$ and observed log price $Y_i(t_{i,j})$,

$$Y_i(t_{i,j}) = X_i(t_{i,j}) + \varepsilon_i(t_{i,j}),$$

$X_i(t)$: true log price of asset i, and microstructure noise $\varepsilon_i(\cdot)$: i.i.d. with zero mean, and independent of $X_i(t)$.

Nonsynchronization: stocks' transactions occur at distinct times and the prices of different stocks are recorded at mismatched time points.


Stock 3
Stock 2
Stock 1
Time

Price Model

$X_t = (X_{1t}, \cdots, X_{pt})^\dagger$: log price of p assets

$$dX_t = \mu_t\, dt + \sigma_t\, dW_t, \qquad t \in [0, 1],$$

where W_t: p-dimensional BM, and σ_t: $p \times p$ matrix.

Integrated volatility matrix:

$$\Gamma = \int_0^1 \gamma(t)\, dt, \qquad \gamma(t) = \sigma_t \sigma_t^\dagger$$

Goal: Estimate Γ based on data $Y_i(t_{ij})$.

Methodology:

1. Form realized volatility (RV) matrix
2. Regularize RV matrix

Realized co-volatility

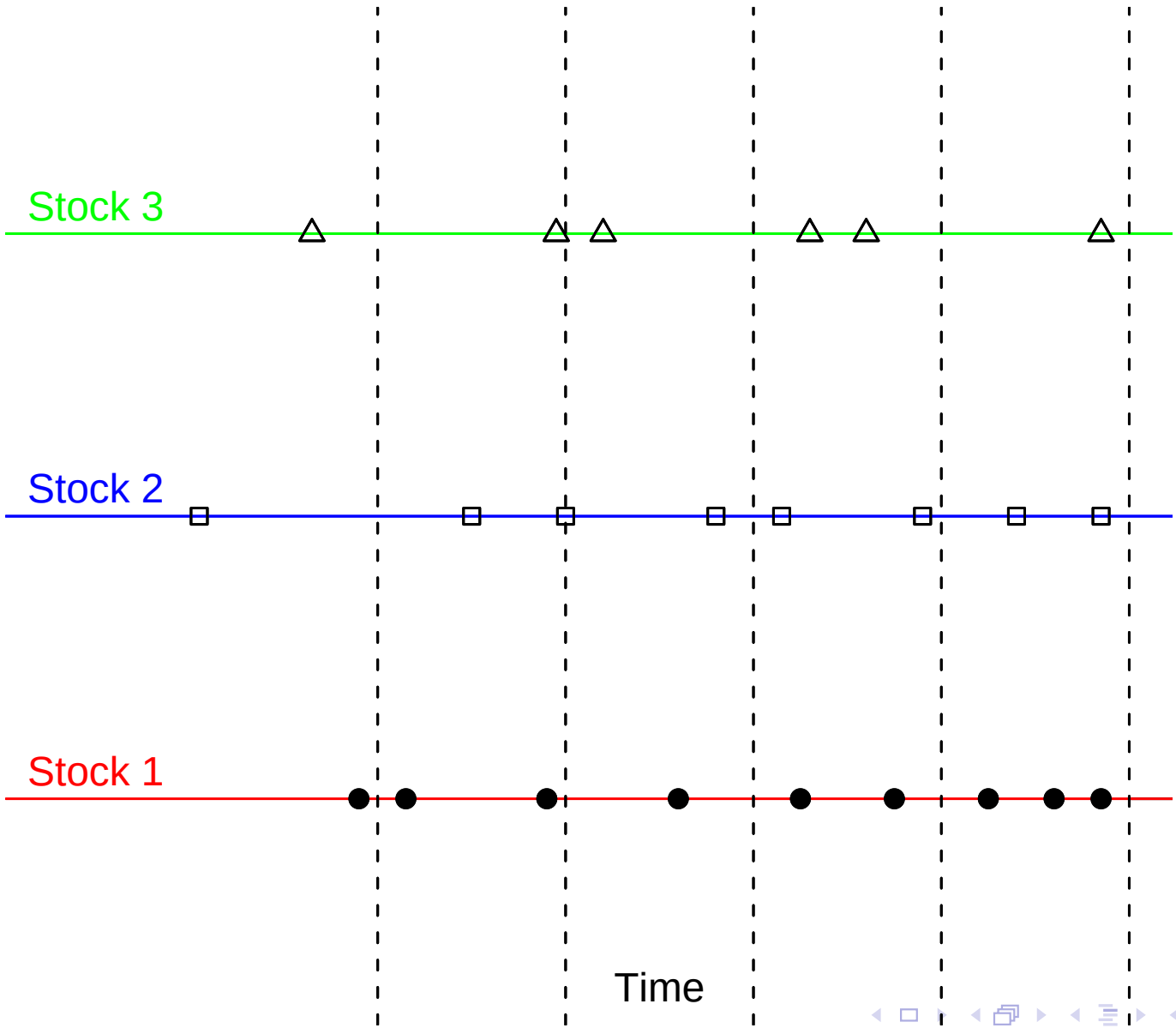
$\tau = \{\tau_r = r/m, r = 1, \cdots, m\}$: pre-determined sampling frequency.

For assets i_1 and i_2, select previous-tick times:

$$\tau_{i_s,r} = \max\{t_{i_s,j} \leq \tau_r, j = 1, \cdots, n_{i_s}\}, \qquad s = 1, 2$$

Realized volatility matrix $\hat{\Gamma}(\tau)$:

$$\hat{\Gamma}_{i_1,i_2}(\tau) = \sum_{r=1}^{m} [Y_{i_1}(\tau_{i_1,r}) - Y_{i_1}(\tau_{i_1,r-1})] [Y_{i_2}(\tau_{i_2,r}) - Y_{i_2}(\tau_{i_2,r-1})].$$


Stock 3
Stock 2
Stock 1
Time

Take $n = (n_1 + \cdots + n_p)/p$,

$$\tau^k = \tau + (k-1)/n, \qquad k = 1, \cdots, K = [n/m]$$

$$\widehat{\Gamma} = \frac{1}{K} \sum_{k=1}^{K} \hat{\Gamma}(\tau^k)$$

where $\widehat{\Gamma}_{ii}$ are adjusted by subtracting them from estimated noise variance components,

$$\frac{2m}{n_i} \sum_{\ell=1}^{n_i} [Y_i(t_{i,\ell}) - Y_i(t_{i,\ell-1})]^2$$

Matrix Size

For each entry $\hat{\Gamma}_{i_1,i_2}$:

$$\hat{\Gamma}_{i_1,i_2} - \Gamma_{i_1,i_2} = O_P(n^{-\eta}), \qquad \eta = 1/2, 1/3, 1/4, 1/6$$

Dimension Reduction For moderate to large p, $(p^2 + p)/2$ entries in Γ: too many parameters and too much random fluctuation.

Issue: Usual dimension reduction techniques are **not applicable** to non-synchronized data.

Numerical Illustration

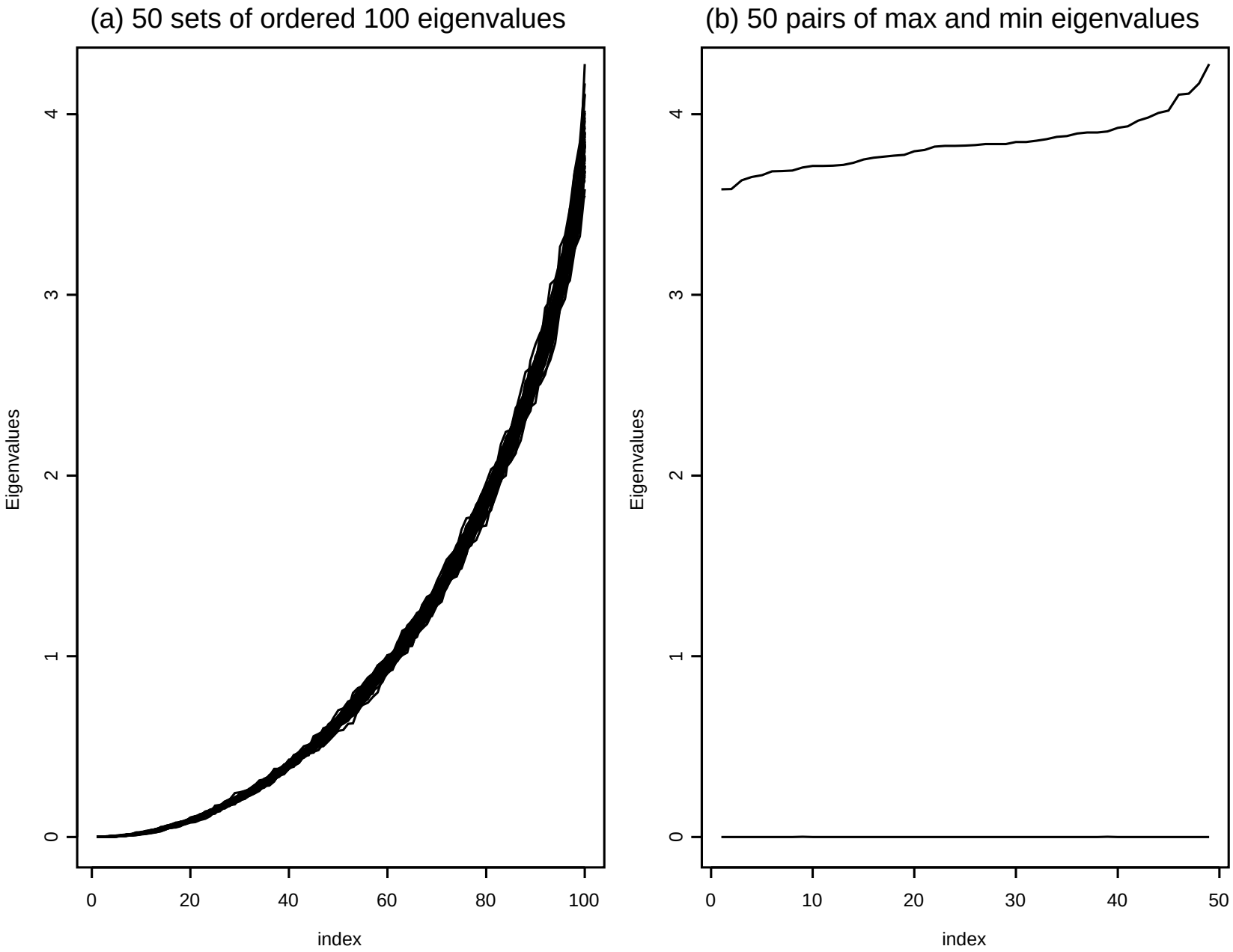
$X(t) = (W_1(t), \cdots, W_p(t))$: vector of p independent Brownian motions.

Observations = $X(k/n)$, $k = 0, 1, \cdots, n$.

$$\Gamma = I_p, \qquad \widehat{\Gamma} = \left(\widehat{\Gamma}_{ij}\right), \qquad \widehat{\Gamma}_{ij} = \frac{1}{n}\sum_{k=1}^{N} Z_{ik} Z_{jk}$$

$$Z_{ik} = \sqrt{n}[W_i(k/n) - W_i((k-1)/n)] \sim N(0,1)$$

Take $n = 100$ and $p = 100$. We compute the eigenvalues of $\widehat{\Gamma}$ in a simulation with 50 replications.


(a) 50 sets of ordered 100 eigenvalues
Eigenvalues
0
1
2
3
4
0
20
40
60
80
100
index
(b) 50 pairs of max and min eigenvalues
Eigenvalues
0
1
2
3
4
0
10
20
30
40
50
index

Regularize Volatility Matrix

Write $\mathbf{\Gamma} = (\Gamma_{ij})$

Sparsity: assume $\mathbf{\Gamma}$ has a sparse representation

$$\sum_{j=1}^{p} |\Gamma_{ij}|^{\delta} \leq M\,\pi(p), \qquad i = 1, \cdots, p, \qquad E[M] \leq C,$$

where $0 \leq \delta < 1$ and $\pi(p) = 1, \log p$, or a small power of p.

Examples: (1) Block diagonal matrix

(2) Matrix with decay elements from diagonal

(3) Matrix with small number of non-zero elements in each row

(4) Random permutations of rows and columns for above matrices

Estimation with Regularization

Write $\hat{\mathbf{\Gamma}} = (\hat{\Gamma}_{ij})$.

Thresholding: for sparse $\mathbf{\Gamma}$, regularize $\hat{\mathbf{\Gamma}}$ by thresholding

$$\mathcal{T}_{\varpi}[\hat{\mathbf{\Gamma}}] = \left(\hat{\Gamma}_{ij} 1(|\hat{\Gamma}_{ij}| \geq \varpi)\right),$$

where ϖ is a threshold.

Asymptotic Theory

Define matrix norm

$$\|\mathbf{\Gamma}\|_2 = \sup\{\|\mathbf{\Gamma}\,\mathbf{x}\|_2, \|\mathbf{x}\|_2 = 1\} = \text{max absolute eigenvalue}$$

Technical conditions

A1: For some $\beta > 0$,

$$\max_{1\le i\le p}\max_{0\le t\le 1} E\left[|\gamma_{ii}(t)|^{\beta}\right] < \infty, \qquad \max_{1\le i\le p}\max_{0\le t\le 1} E\left[|\mu_i(t)|^{\beta}\right] < \infty,$$

$$\max_{1\le i\le p} E\left[|\varepsilon_i(t_{i\ell})|^{2\,\beta}\right] < \infty.$$

A2: Each asset has at least one observation between consecutive time points of the selected sampling frequency. With $n = (n_1 + \cdots + n_p)/p$,

$$C_1 \le \min_{1\le i\le p}\frac{n_i}{n} \le \max_{1\le i\le p}\frac{n_i}{n} \le C_2, \qquad \max_{1\le i\le p}\max_{1\le \ell\le n_i} |t_{i\ell} - t_{i,\ell-1}| = O(n^{-1}).$$

Theorem For sparse $\mathbf{\Gamma}$, under conditions A1-A2, we have

$$\|\mathcal{T}_{\varpi}[\hat{\mathbf{\Gamma}}] - \mathbf{\Gamma}\|_2 = O_P\left(\pi(p)\,\left[p\, e_n^{\beta/2}\right]^{2(1-\delta)/\beta}\right),$$

where $\varpi = e_n\, p^{2/\beta}\, \log\log(n \wedge p)$, and for noisy data, $e_n \sim n^{-1/6}$,

$$\textbf{convergence rate} = \pi(p)\,\left[p\, n^{-\beta/12}\right]^{2(1-\delta)/\beta}$$

for noiseless data, $e_n \sim n^{-1/3}$,

$$\textbf{convergence rate} = \pi(p)\,\left[p\, n^{-\beta/6}\right]^{2(1-\delta)/\beta}$$

Some Insights

(1) Multiple random sources;

(2) Fat tails;

(3) Non-synchronization problem.

$$Y_i(t_{i,j}) = X_i(t_{i,j}) + \varepsilon_i(t_{i,j})$$

$$X_i(t_{i,j}) = \int_0^{t_{i,j}} \mu_{i,s}\, ds + \int_0^{\mathbf{t_{i,j}}} \sigma_{\mathbf{i,s}}\, \mathbf{dW_s}$$

Simulations

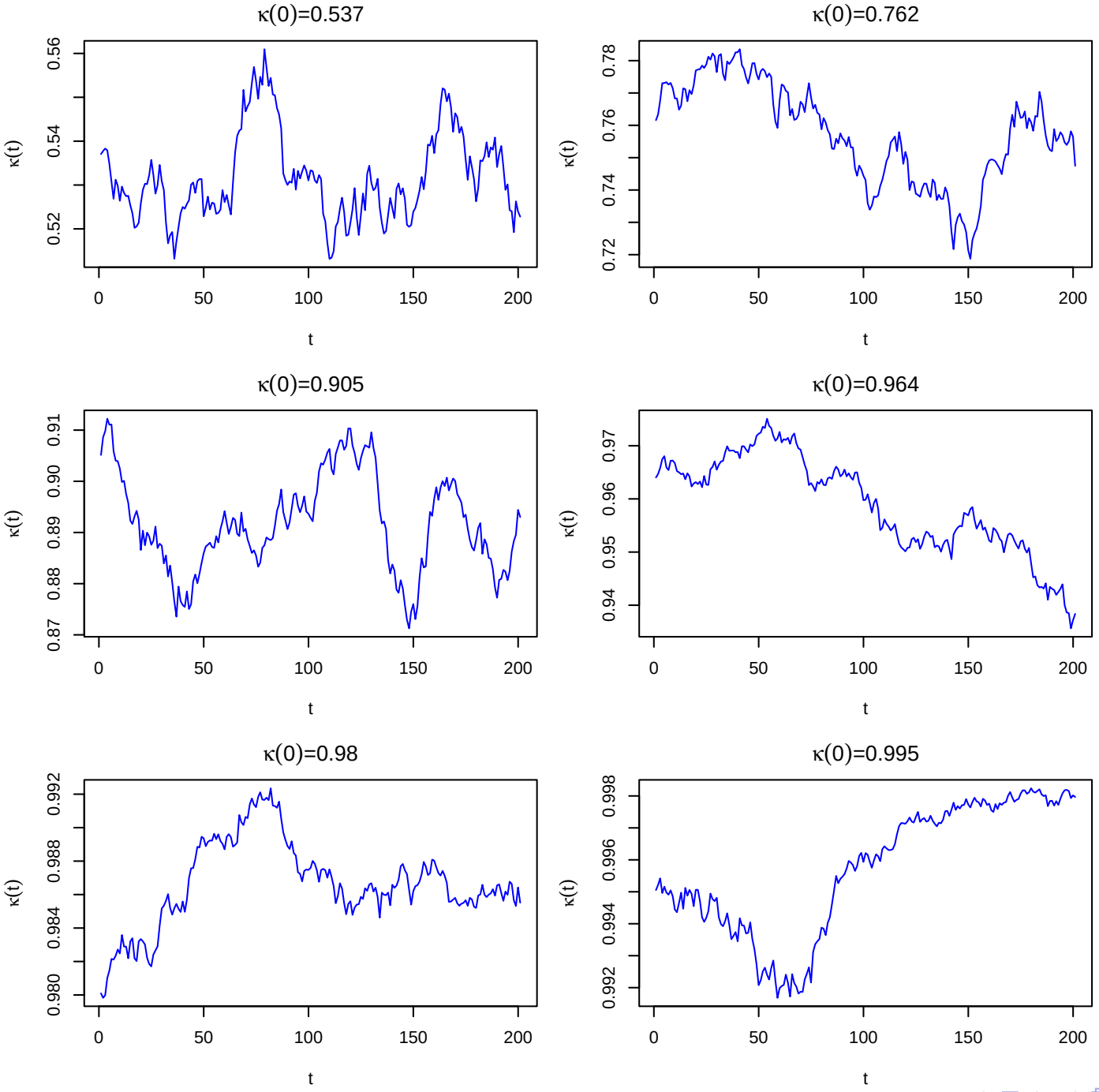
$$Y_i(t_{ij}) = \int_0^{t_{ij}} \sigma_{i,s}\, dB_s + \varepsilon_i(t_{ij}), \quad i = 1, \cdots, p, \quad j = 1, \cdots, n$$

$p = 512$, $n = 200$, $\gamma(t) = \sigma_t\, \sigma_t^{\dagger}$. $\gamma_{ii}(t)$: geometric OU, sum of two CIR, Nelson GARCH diffusion, and two-factor log linear SV.

$$\gamma_{ij}(t) = \{\kappa(t)\}^{|i-j|} \sqrt{\gamma_{ii}(t)\gamma_{jj}(t)}, \qquad 1 \le i \ne j \le p,$$

$$\kappa(t) = \frac{e^{2\,u(t)} - 1}{e^{2\,u(t)} + 1}, \qquad du(t) = 0.03\,[0.64 - u(t)]\,dt + 0.118\,u(t)\,dW_{\kappa,t},$$

$$W_{\kappa,t} = \sqrt{0.96}\, W^0_{\kappa,t} - 0.2 \sum_{i=1}^{p} B_{it}/\sqrt{p},$$


κ(0)=0.537
κ(0)=0.762
κ(0)=0.905
κ(0)=0.964
κ(0)=0.98
κ(0)=0.995
κ(t)
t
0
50
100
150
200
0.52
0.54
0.56
0.72
0.74
0.76
0.78
0.87
0.88
0.89
0.90
0.91
0.94
0.95
0.96
0.97
0.980
0.984
0.988
0.992
0.992
0.994
0.996
0.998

Simulate γ_{ii}: Brownian motions governing γ_{ii} are constructed as follows

$$W_{it}^1 = \rho_i B_{it} + \sqrt{1-\rho_i^2}\, U_{it}^1, \qquad W_{it}^2 = \rho_i B_{it} + \sqrt{1-\rho_i^2}\, U_{it}^2,$$

where U_i^1 and U_i^2 are independent standard BW, and

$$\rho_i = \begin{cases} -0.62, & 1 \le i \le p/4, \\ -0.50, & p/4 < i \le p/2, \\ -0.25, & p/2 < i \le 3p/4, \\ -0.30, & 3p/4 < i \le p. \end{cases}$$

Geometric OU process

$$d \log \gamma_{ii}(t) = -0.6\,(0.157 + \log \gamma_{ii}(t))\,dt + 0.25\,dW_{it}^{1}$$

Sum of two CIR:

$$\gamma_{ii}(t) = 0.98\,(v_{1,t} + v_{2,t})$$

$$dv_{1,t} = 0.0429\,(0.108 - v_{1,t})\,dt + 0.1539\,\sqrt{v_{1,t}}\,dW_{i,t}^{1}$$

$$dv_{2,t} = 3.74\,(0.401 - v_{2,t})\,dt + 1.4369\,\sqrt{v_{2,t}}\,dW_{i,t}^{2}$$

Nelson GARCH diffusion:

$$d\gamma_{ii}(t) = [0.1 - \gamma_{ii}(t)]\,dt + 0.2\,\gamma_{ii}(t)\,dW_{i,t}^{1}$$

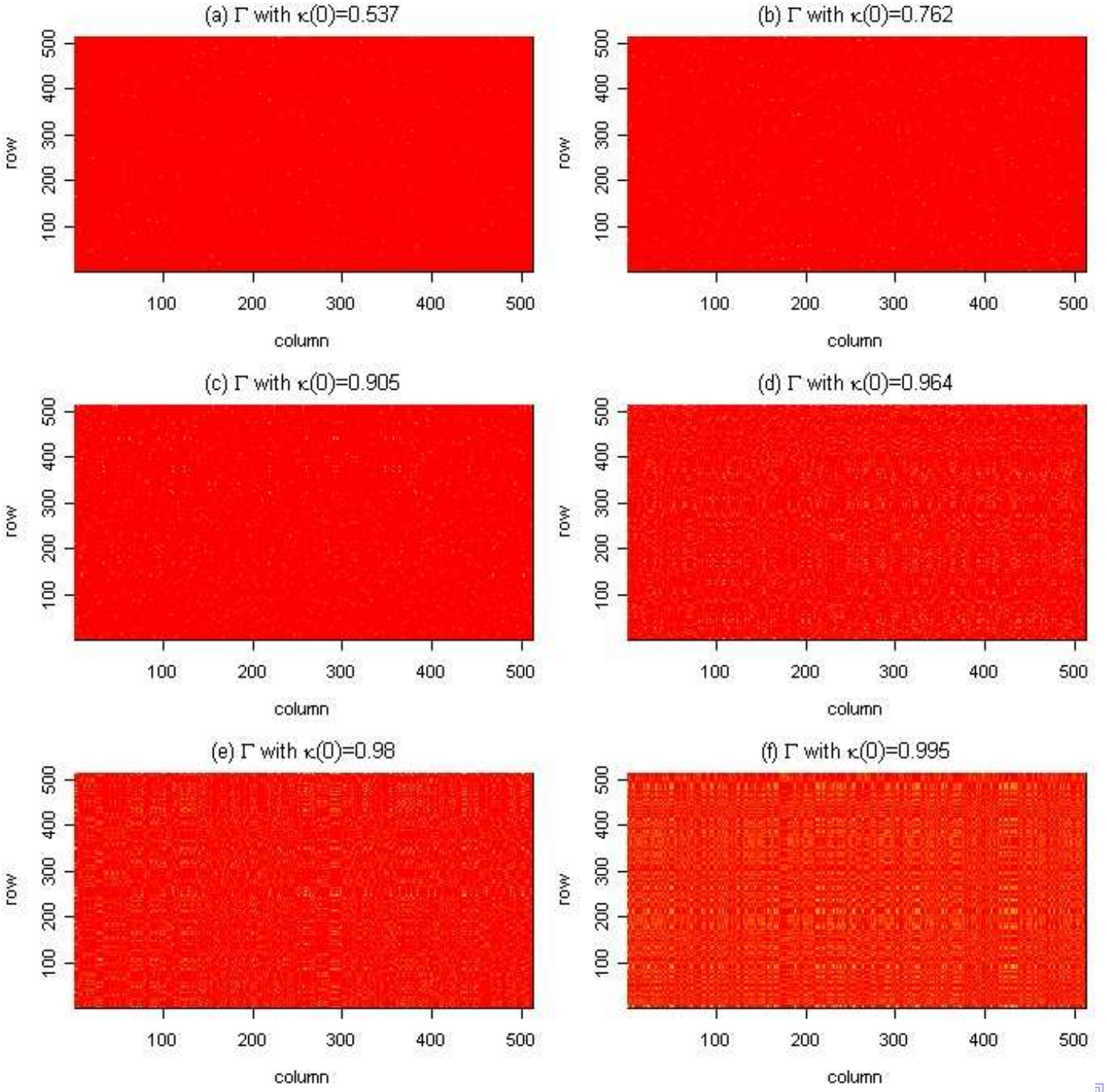
Two-factor log linear SV model:

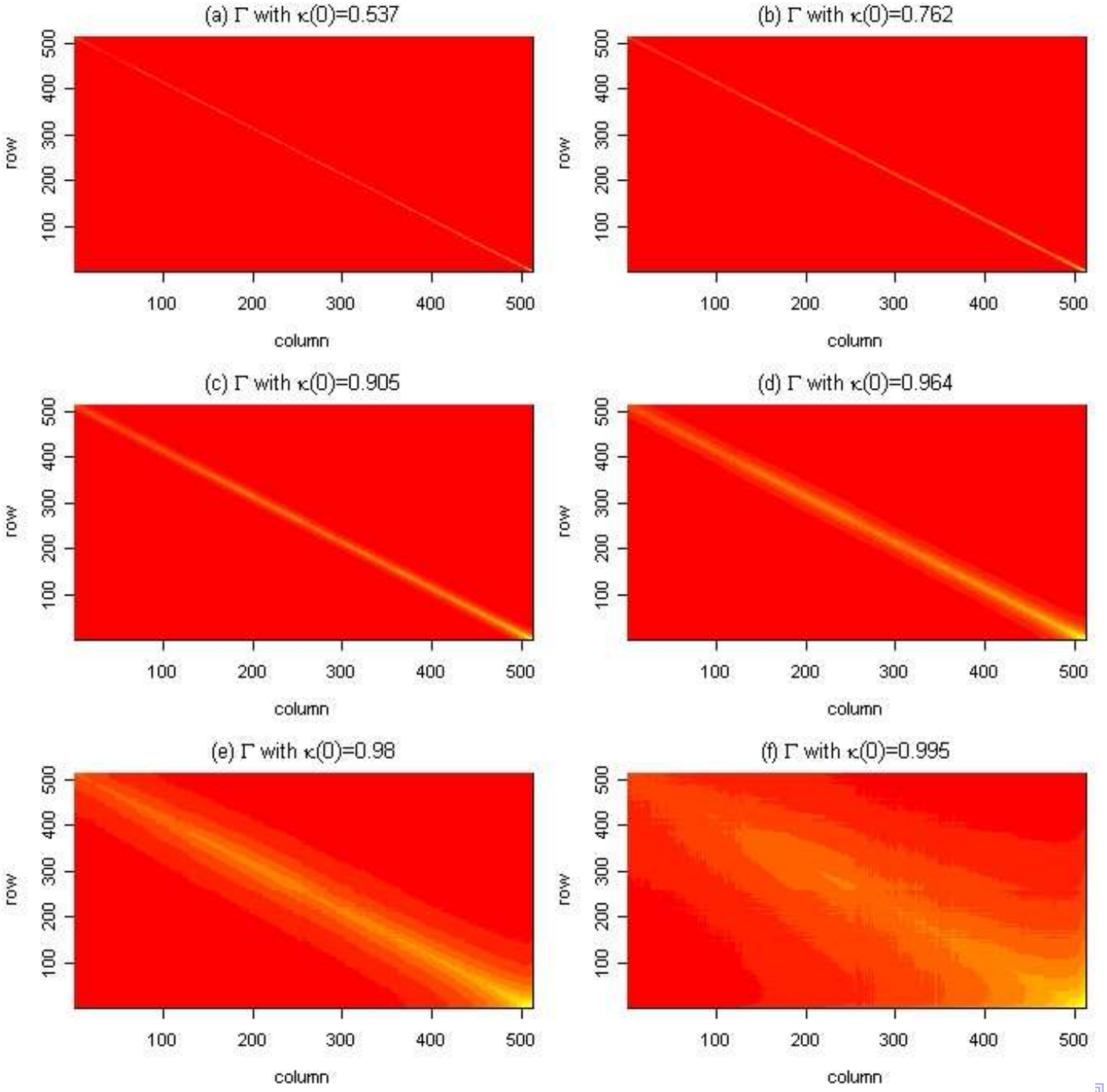
$$\gamma_{ii}(t) = e^{-6.8753}\,\text{s-exp}(0.04\,v_{1,t} + 1.5\,v_{2,t} - 1.2)$$

$$d\,v_{1,t} = -0.00137\,v_{1,t}\,dt + dW^1_{i,t}$$

$$d\,v_{2,t} = -1.386\,v_{2,t}\,dt + (1 + 0.25\,v_{2,t})\,dW^2_{i,t}$$

$$\text{s-exp}(u) = \begin{cases} e^u, & \text{if } u \le \log(8.5), \\ 8.5\,\{1 - \log(8.5) + u^2/\log(8.5)\}^{1/2}, & \text{if } u > \log(8.5). \end{cases}$$


(a) Γ with κ(0)=0.537
(b) Γ with κ(0)=0.762
(c) Γ with κ(0)=0.905
(d) Γ with κ(0)=0.964
(e) Γ with κ(0)=0.98
(f) Γ with κ(0)=0.995
row
column
100
200
300
400
500


(a) Γ with κ(0)=0.537
(b) Γ with κ(0)=0.762
(c) Γ with κ(0)=0.905
(d) Γ with κ(0)=0.964
(e) Γ with κ(0)=0.98
(f) Γ with κ(0)=0.995
row
column
100
200
300
400
500

$\varepsilon_i(t_{ij})$ i.i.d $N(0, a^2)$ with $a = 0.002, 0.127, 0.2$ for low, medium and high noise levels.

MSE for noisy synchronized data

		$\kappa(0)$				
Noise	Estimator	0.537	0.762	0.905	0.964	0.995
Low	$\hat{\mathbf{\Gamma}}$	5.595	6.039	7.511	9.959	18.270
Low	$\mathcal{T}_\varpi[\hat{\mathbf{\Gamma}}]$	0.845	2.456	4.595	7.457	
Med	$\hat{\mathbf{\Gamma}}$	5.641	6.097	7.649	10.479	18.398
Med	$\mathcal{T}_\varpi[\hat{\mathbf{\Gamma}}]$	0.871	2.466	4.101	7.680	
High	$\hat{\mathbf{\Gamma}}$	5.769	6.234	7.717	10.521	19.26
High	$\mathcal{T}_\varpi[\hat{\mathbf{\Gamma}}]$	0.896	2.429	4.043	8.765	

Simulate $3n = 600$ observations and divide into 200 groups of 3 observations. Randomly pick up one from each group to form non-synchronized data.

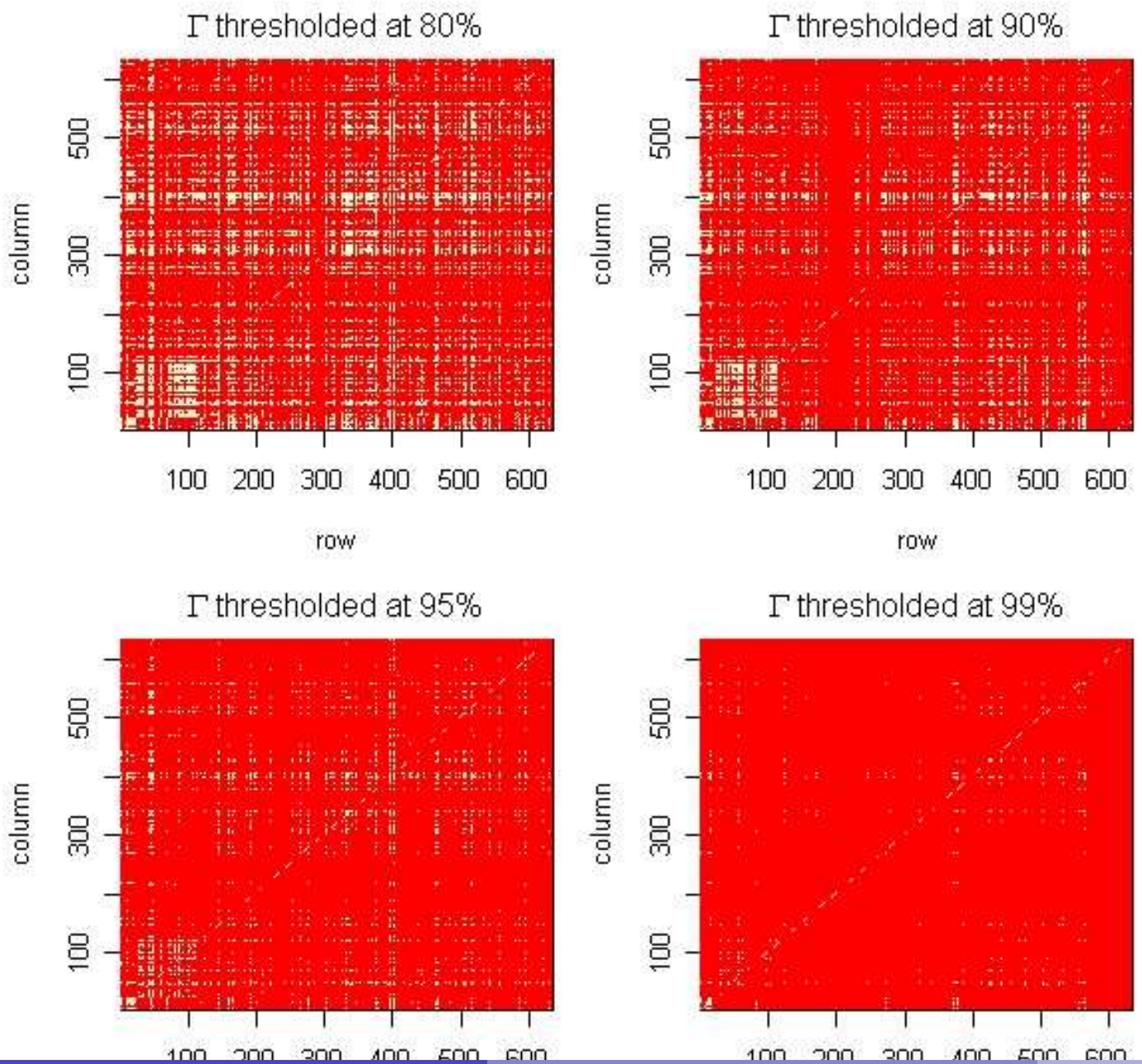
MSE for noisy non-synchronized data

		$\kappa(0)$				
Noise	Estimator	0.537	0.762	0.905	0.964	0.995
Low	$\hat{\mathbf{\Gamma}}$	12.86	16.99	27.13	48.37	152.75
Low	$\mathcal{T}_{\varpi}[\hat{\mathbf{\Gamma}}]$	3.842	5.281	13.38	30.29	121.15
Med	$\hat{\mathbf{\Gamma}}$	12.98	17.10	27.15	48.57	153.57
Med	$\mathcal{T}_{\varpi}[\hat{\mathbf{\Gamma}}]$	3.374	4.728	11.662	30.53	123.22
High	$\hat{\mathbf{\Gamma}}$	13.16	17.15	27.50	48.07	151.85
High	$\mathcal{T}_{\varpi}[\hat{\mathbf{\Gamma}}]$	3.997	4.902	11.70	29.98	100.13

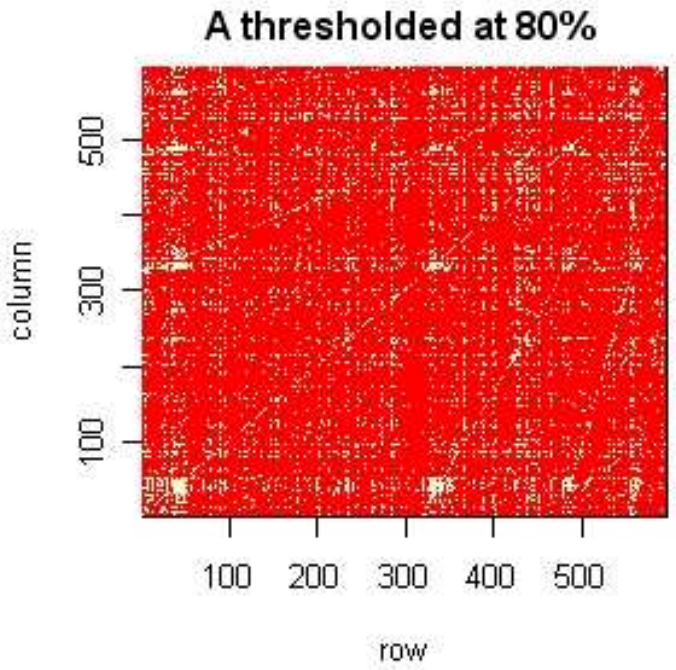
Application

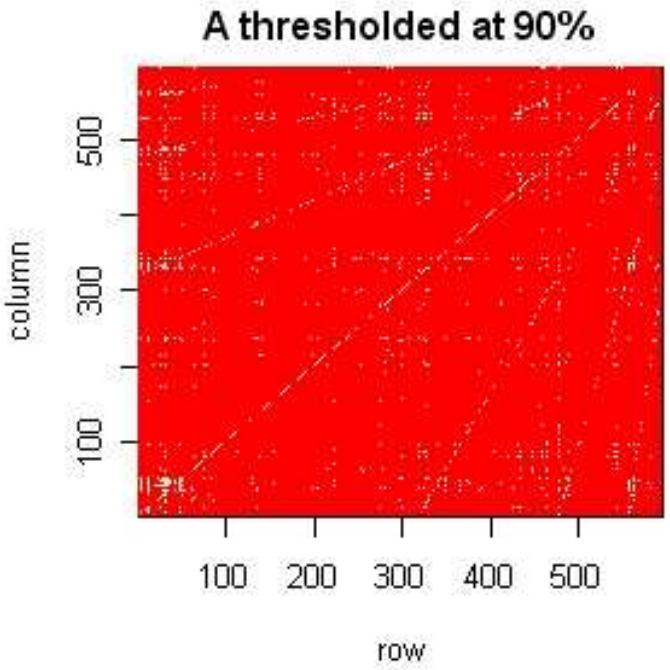
High-frequency price data on 630 stocks traded in Shanghai market over 177 days in 2003. For each day we compute volatility matrix estimator $\tilde{\Gamma}_i$, $i = 1, \cdots, 177$.

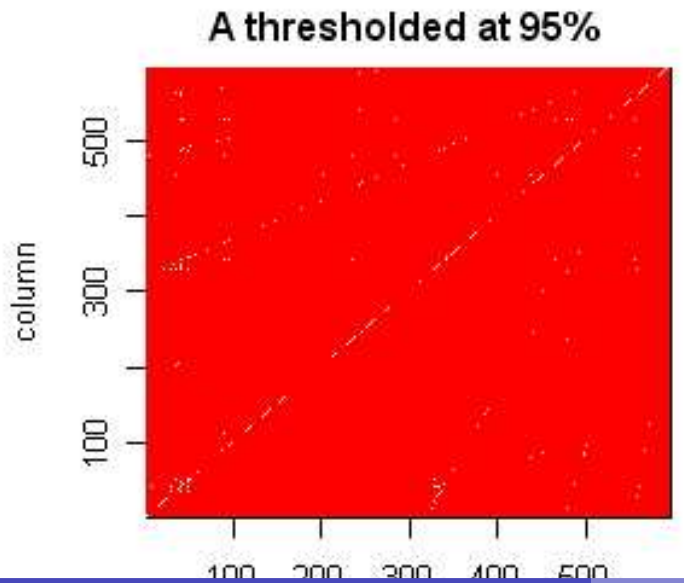
Average RV matrix at various threshold levels

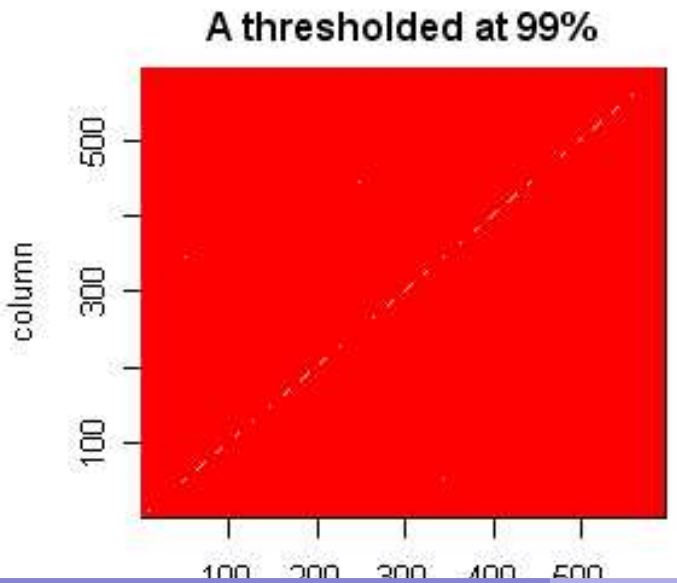


Average wavelet RV matrix at various threshold levels

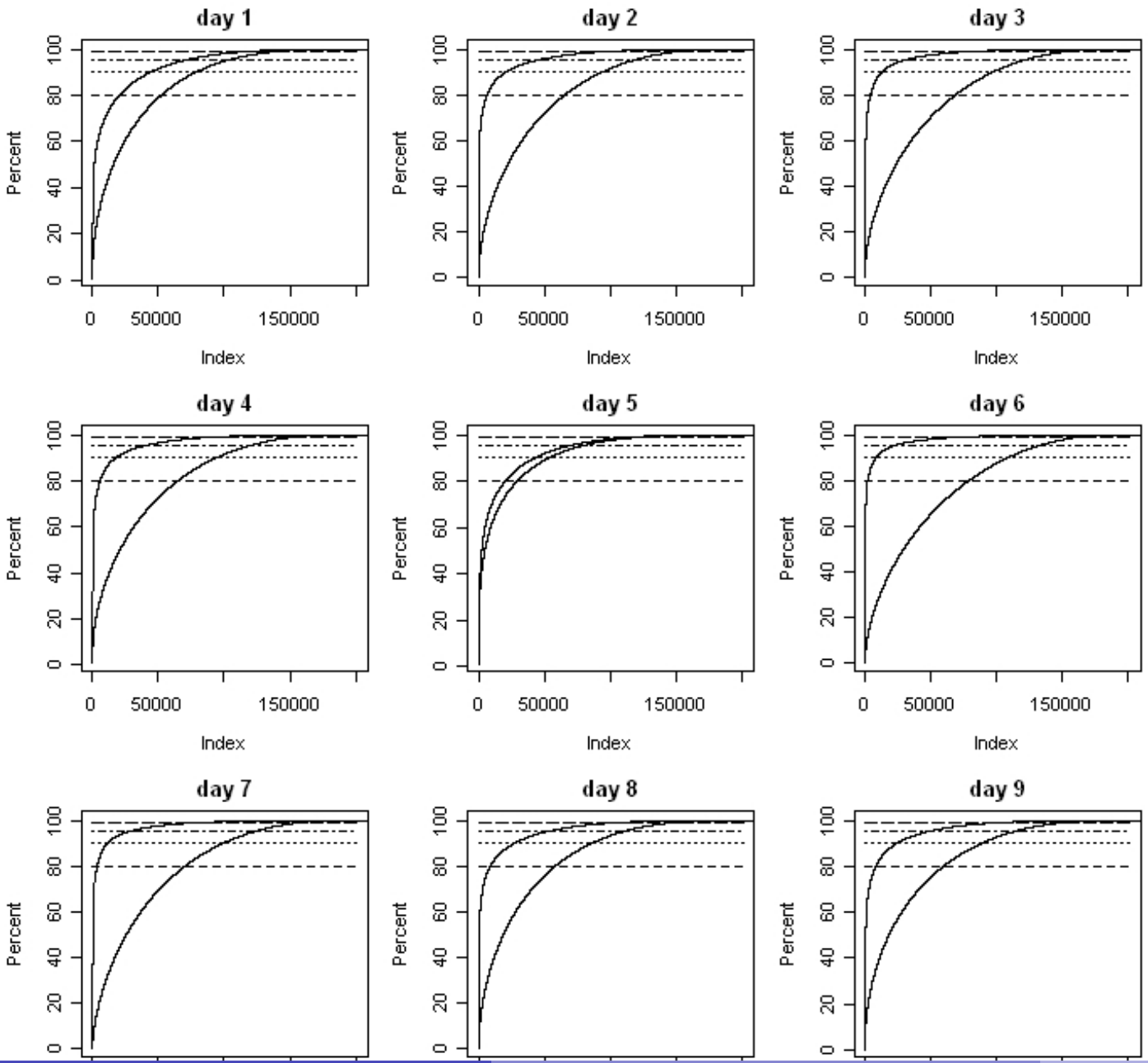








Concentration plots of RV and wavelet RV estimators for 9 days



Threshold $\tilde{\Gamma}_i$

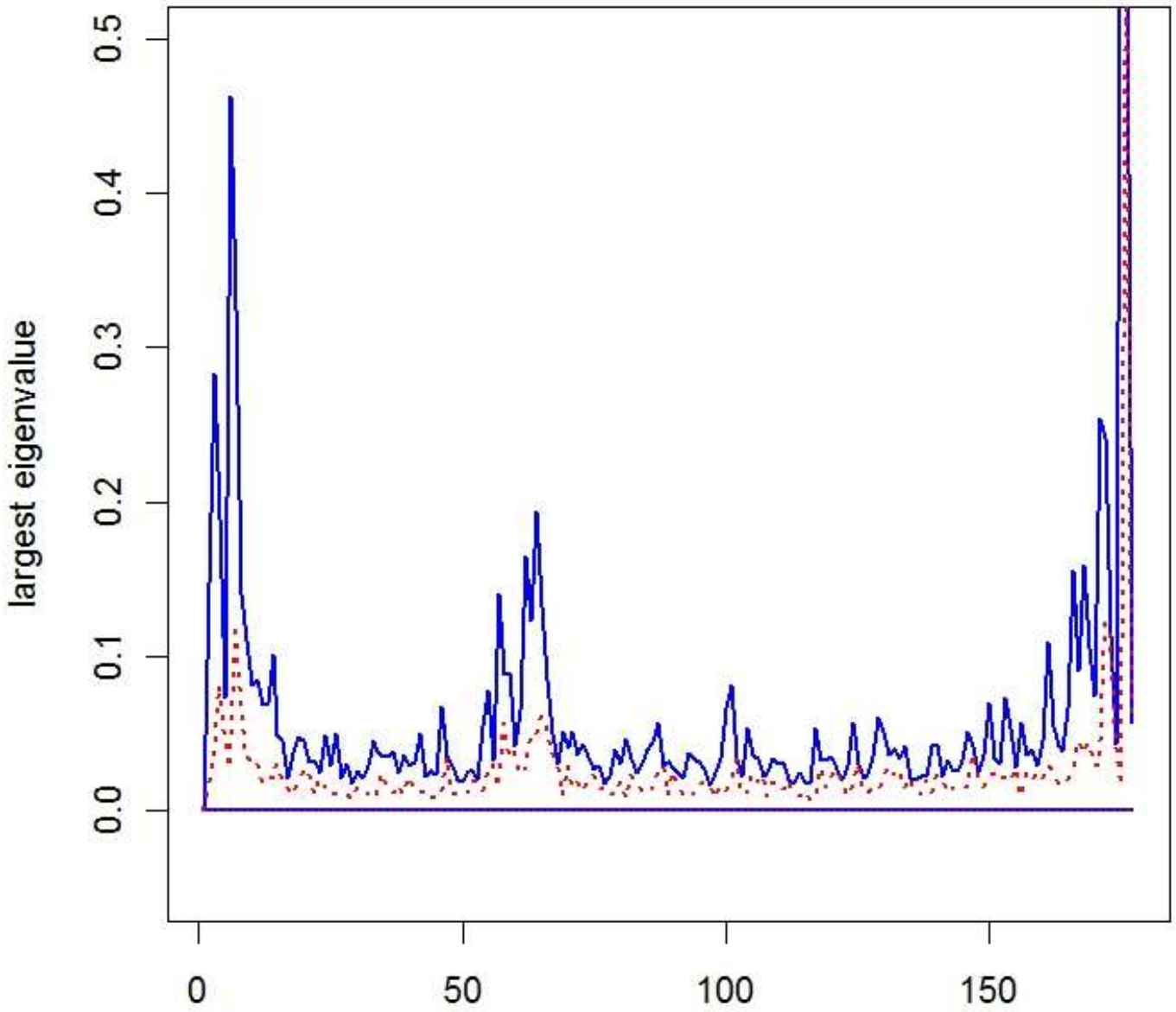
Calibrate threshold value $\varpi_{i,a}$= a-quantile of the absolute entries of $\tilde{\Gamma}_i$ and choose the value of a to minimize

$$\Lambda(a) = \sum_{i=1}^{176} \left\| \tilde{\Gamma}_{i+1} - \mathcal{T}_{\varpi_{i,a}} \left[\tilde{\Gamma}_i \right] \right\|_2^2$$

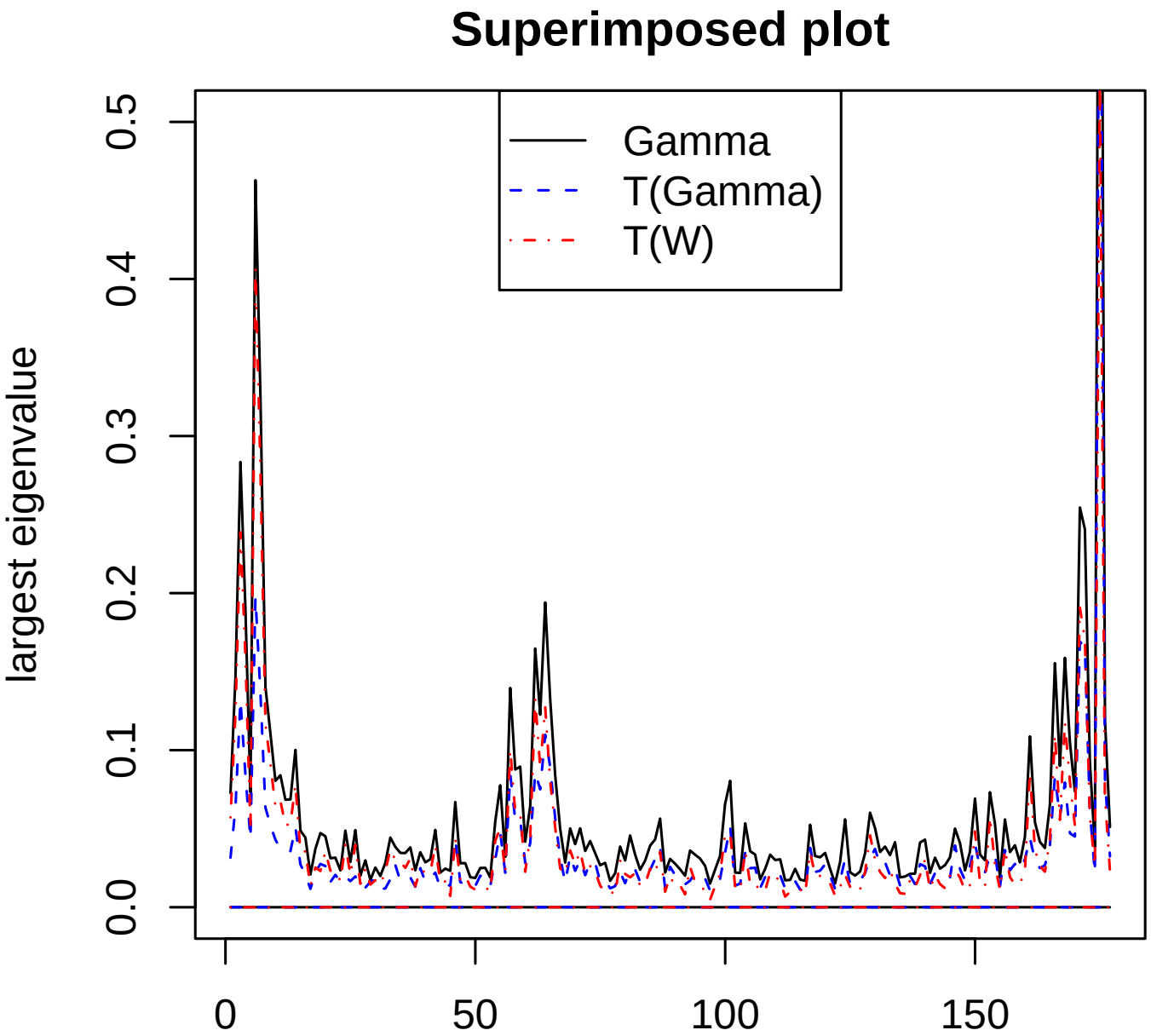
Find $\hat{a} = 0.95$ and evaluate $\mathcal{T}_{\varpi_{i,0.95}} \left[\tilde{\Gamma}_i \right]$.

Compute the eigenvalues for $\tilde{\Gamma}_i$ and $\mathcal{T}_{\varpi_{i,0.95}} \left[\tilde{\Gamma}_i \right]$ and plot their corresponding largest eigenvalues for comparison.

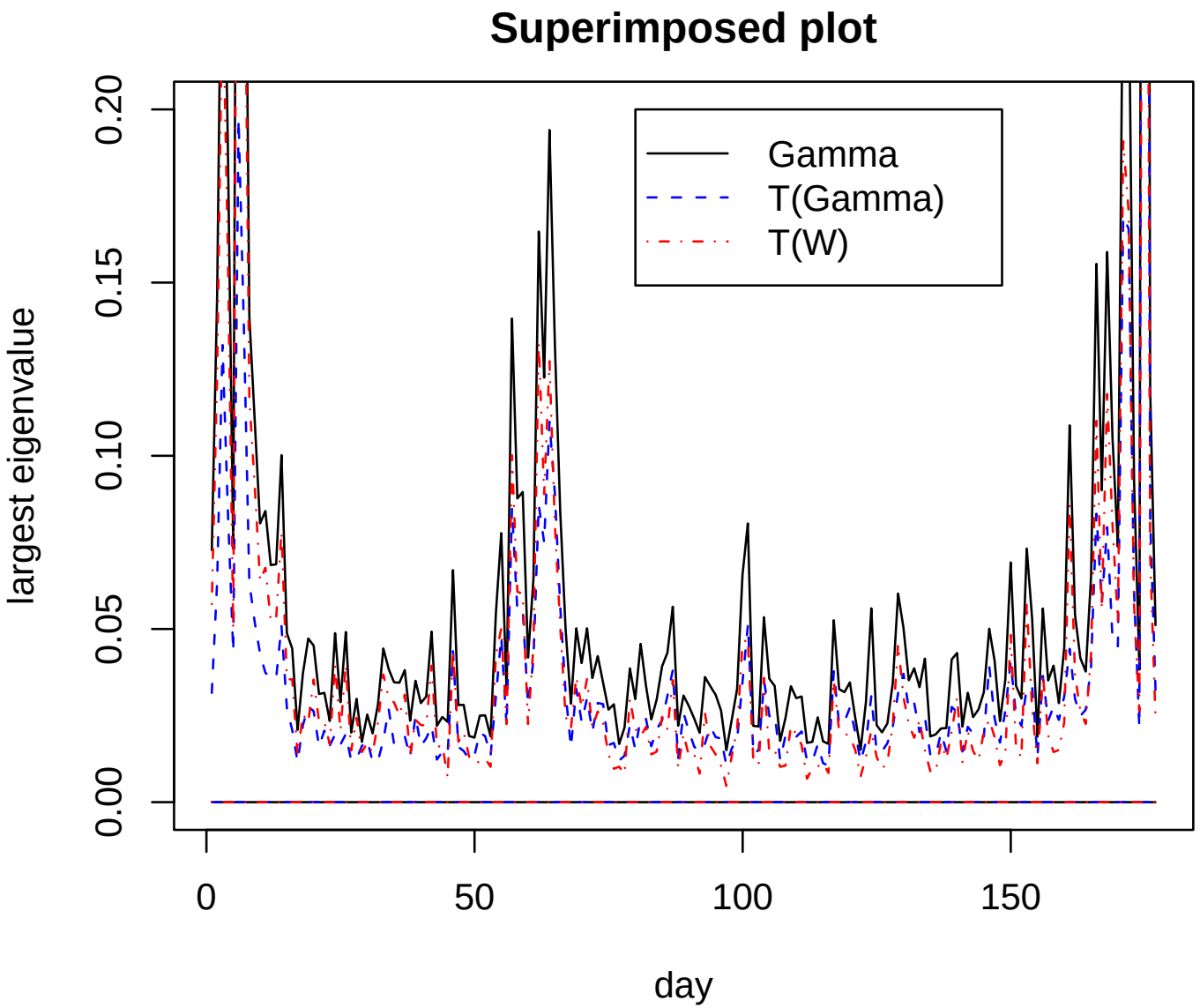
Largest eigenvalues of $\tilde{\Gamma}_i$ and their regularizations



Largest eigenvalues for $\tilde{\Gamma}_j$, regularizations, and wavelet versions



Zoom in plot


Superimposed plot
Gamma
T(Gamma)
T(W)
largest eigenvalue
0.00
0.05
0.10
0.15
0.20
0
50
100
150
day

Concluding Remarks

1. Good matrix estimators may perform poorly when the matrix size is very large. We need to regularize large sample covariance and RV matrix estimators.

2. For sparse matrices, thresholding yields good performance for sample covariance and RV based matrix estimators.

Papers can be down-loaded from

http://www.stat.uconn.edu/~yzwang