Spectral distributions of high-dimensional sample correlation matrices under infinite variance

Johannes Heiny

Ruhr-University Bochum

Joint work with
Jianfeng Yao (HKU),
Thomas Mikosch and Jorge Yslas (Copenhagen).

Random Matrices and Complex Data Analysis Workshop,
December 10-12, 2019, Shanghai

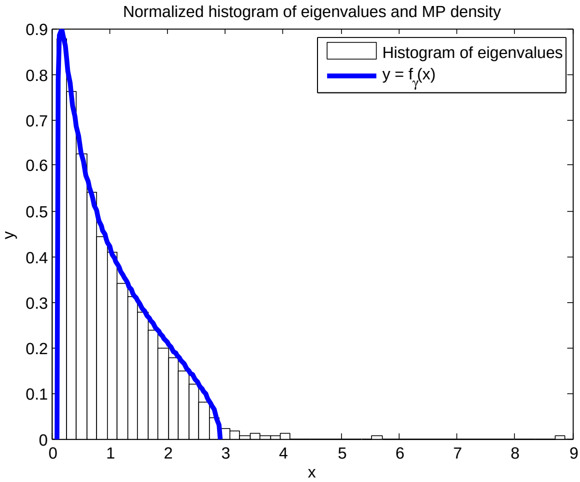


Figure: These are NOT spikes!

Setup for the picture

- **Data matrix** $\boldsymbol{X} = \boldsymbol{X}_n$: $p \times n$ matrix with iid centered entries and generic variable $X \stackrel{\mathrm{d}}{=} X_{11}$.

$$\boldsymbol{X} = (X_{it})_{i=1,\ldots,p;t=1,\ldots,n}$$

- **Sample covariance matrix** $\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}'$
- **Ordered eigenvalues** of $\boldsymbol{S}$

$$\lambda_1(\boldsymbol{S}) \geq \lambda_2(\boldsymbol{S}) \geq \cdots \geq \lambda_p(\boldsymbol{S})$$

- **Sample correlation matrix**

$$\mathbf{R} = (\mathrm{diag}(\boldsymbol{S}))^{-1/2}\boldsymbol{S}\,(\mathrm{diag}(\boldsymbol{S}))^{-1/2}\,.$$

Regular variation

- **Regular variation** with index $\alpha > 0$:

$$\mathbb{P}(|X| > x) = x^{-\alpha} L(x),$$

where L is a slowly varying function.

- This implies $\mathbb{E}[|X|^{\alpha+\varepsilon}] = \infty$ for any $\varepsilon > 0$.
- **Normalizing sequence** (a_{np}^2) such that

$$np\, \mathbb{P}(X^2 > a_{np}^2 x) \to x^{-\alpha/2}, \quad \text{as } n \to \infty \text{ for } x > 0.$$

Then $a_{np} = (np)^{1/\alpha} \ell(np)$ for a slowly varying function ℓ.

Reduction to Diagonal

Diagonal

$\boldsymbol{X}$ with iid regularly varying entries $\alpha \in (0,4)$ and $p = n^\beta$ with $\beta \in [0,1]$. We have

$$a_{np}^{-2} \|\boldsymbol{X}\boldsymbol{X}' - \operatorname{diag}(\boldsymbol{X}\boldsymbol{X}')\| \stackrel{\mathbb{P}}{\to} 0\,,$$

where $\|\cdot\|$ denotes the spectral norm.

$$(\boldsymbol{X}\boldsymbol{X}')_{ij} = \sum_{t=1}^{n} X_{it}X_{jt}.$$

Eigenvalues

- **Weyl's inequality**

$$\max_{i=1,\ldots,p} \big|\lambda_i(\mathbf{A}+\mathbf{B}) - \lambda_i(\mathbf{A})\big| \le \|\mathbf{B}\|\,.$$

- Choose $\mathbf{A}+\mathbf{B} = \boldsymbol{X}\boldsymbol{X}'$ and $\mathbf{A} = \mathrm{diag}(\boldsymbol{X}\boldsymbol{X}')$ to obtain

$$a_{np}^{-2} \max_{i=1,\ldots,p} \big|\lambda_i(\boldsymbol{X}\boldsymbol{X}') - \lambda_i(\mathrm{diag}(\boldsymbol{X}\boldsymbol{X}'))\big| \stackrel{\mathbb{P}}{\to} 0\,, \quad n \to \infty\,.$$

- **Note:** Limit theory for $(\lambda_i(\boldsymbol{S}))$ reduced to (S_{ii}).

Theorem (Heiny and Mikosch, 2016)

$\boldsymbol{X}$ with iid regularly varying entries $\alpha \in (0,4)$ and $p_n = n^\beta \ell(n)$ with $\beta \in [0,1]$.

1. If $\beta \in [0,1]$, then

$$\boxed{a_{np}^{-2} \max_{i=1,\dots,p} \big|\lambda_i(\boldsymbol{X}\boldsymbol{X}') - \lambda_i(\mathrm{diag}(\boldsymbol{X}\boldsymbol{X}'))\big| \stackrel{\mathbb{P}}{\to} 0\,.}$$

2. If $\beta \in ((\alpha/2-1)_+, 1]$, then

$$a_{np}^{-2} \max_{i=1,\dots,p} \big|\lambda_i(\boldsymbol{X}\boldsymbol{X}') - X^2_{(i),np}\big| \stackrel{\mathbb{P}}{\to} 0\,.$$

Example: Eigenvalues

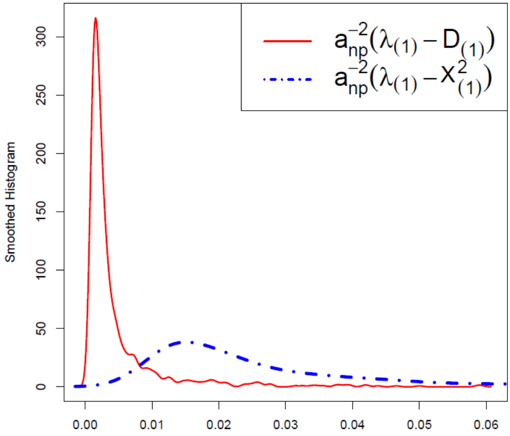


Figure: Smoothed histogram based on 20000 simulations of the approximation error for the normalized eigenvalue $a_{np}^{-2}\lambda_1(\boldsymbol{S})$ for entries X_{it} with $\alpha = 1.6$, $\beta = 1$, $n = 1000$ and $p = 200$.

Eigenvectors

- $\mathbf{v}_k$ unit eigenvector of $\boldsymbol{S}$ associated to $\lambda_k(\boldsymbol{S})$
- Unit eigenvectors of $\mathrm{diag}(\boldsymbol{S})$ are canonical basisvectors $\mathbf{e}_j$.

Eigenvectors

$\boldsymbol{X}$ with iid regularly varying entries with index $\alpha \in (0,4)$ and $p_n = n^\beta \ell(n)$ with $\beta \in [0,1]$. Then for any fixed $k \geq 1$,

$$\|\mathbf{v}_k - \mathbf{e}_{L_k}\|_{\ell_2} \xrightarrow{\mathbb{P}} 0\,, \quad n \to \infty\,.$$

Localization vs. Delocalization

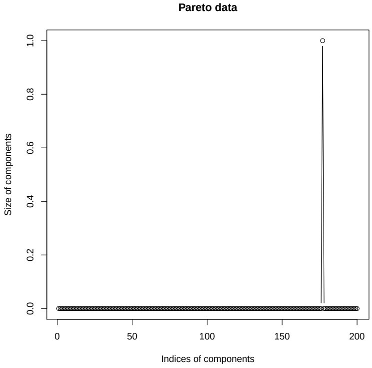


Figure: $X \sim$ Pareto(0.8)

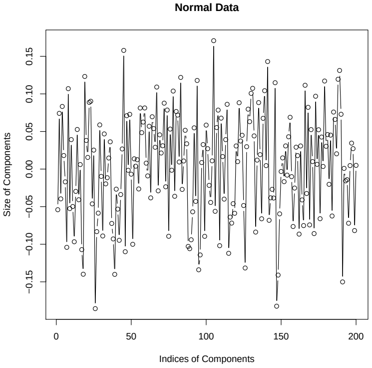


Figure: $X \sim N(0,1)$

Components of eigenvector $\mathbf{v}_1$. $p = 200$, $n = 1000$.

Point process convergence

$$N_n = \sum_{i=1}^{p} \delta_{a_{np}^{-2}\lambda_i(\boldsymbol{X}\boldsymbol{X}')} \stackrel{\mathrm{d}}{\to} \sum_{i=1}^{\infty} \delta_{\Gamma_i^{-2/\alpha}} = N$$

The limit is a PRM on $(0,\infty)$ with mean measure $\mu(x,\infty) = x^{-\alpha/2}, x > 0$, and

$$\Gamma_i = E_1 + \cdots + E_i\,, \quad (E_i) \text{ iid standard exponential.}$$

Point Process of Normalized Eigenvalues

- **Limiting distribution:** For $k \geq 1$,

$$\lim_{n\to\infty} \mathbb{P}(a_{np}^{-2}\lambda_k \leq x) = \lim_{n\to\infty} \mathbb{P}(N_n(x,\infty) < k) = \mathbb{P}(N(x,\infty) < k)$$
$$= \sum_{s=0}^{k-1} \frac{\left(x^{-\alpha/2}\right)^s}{s!}\, \mathrm{e}^{-x^{-\alpha/2}}, \quad x > 0\,.$$

Point Process of Normalized Eigenvalues

- **Limiting distribution:** For $k \ge 1$,

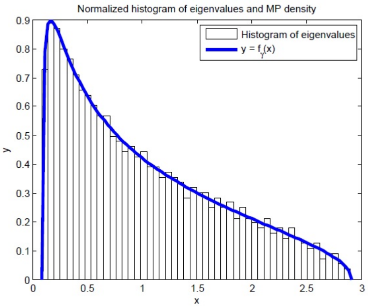
$$\lim_{n\to\infty} \mathbb{P}(a_{np}^{-2}\lambda_k \le x) = \lim_{n\to\infty} \mathbb{P}(N_n(x,\infty) < k) = \mathbb{P}(N(x,\infty) < k)$$
$$= \sum_{s=0}^{k-1} \frac{\left(x^{-\alpha/2}\right)^s}{s!} \mathrm{e}^{-x^{-\alpha/2}}, \quad x > 0\,.$$

- **Largest eigenvalue**

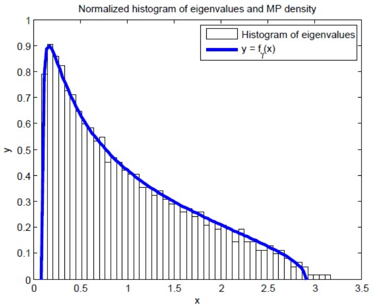
$$\frac{n}{a_{np}^2}\lambda_1(\boldsymbol{S}) \xrightarrow{\mathrm{d}} \Gamma_1^{-\alpha/2}\,,$$

where the limit has a *Fréchet distribution* with parameter $\alpha/2$.
Soshnikov (2006), Auffinger et al. (2009), Auffinger and Tang (2016), Davis et al. $(2014, 2016^2)$, JH and Mikosch (2016)

$\alpha = 3.99$



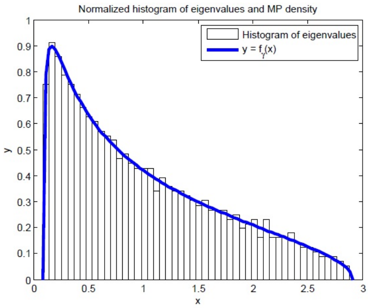
(a) Sample correlation



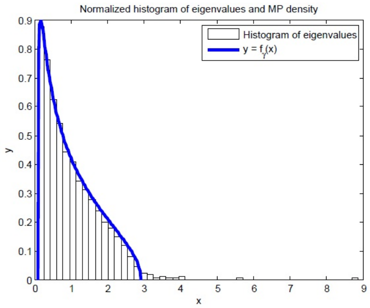
(b) Sample covariance

$$\alpha = 3.99, n = 2000, p = 1000$$

$\alpha = 3$



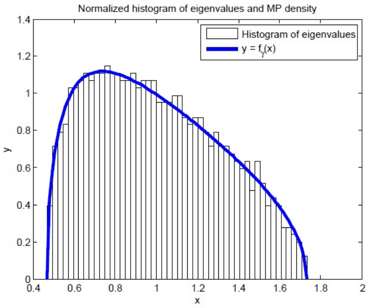
(a) Sample correlation



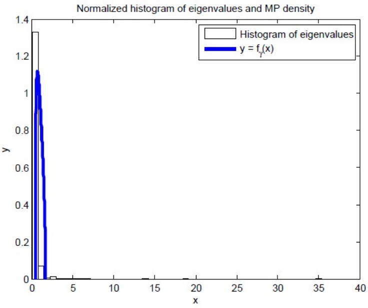
(b) Sample covariance

$$\alpha = 3, n = 2000, p = 1000$$

$\alpha = 2.1$



(a) Sample correlation



(b) Sample covariance

$$\alpha = 2.1, n = 10000, p = 1000$$

Infinite variance, $\alpha < 2$

Limiting spectral distribution of $(\boldsymbol{X}\boldsymbol{X}')$ under $\mathbb{E}[X^2] = \infty$:

- Regular variation with $\alpha < 2$:

$$F_{a_{n+p}^{-2}\boldsymbol{X}\boldsymbol{X}'} \to G_\alpha^\gamma \text{ weakly,}$$

whose density g_α^γ satisfies

$$g_\alpha^\gamma(x) \sim c\, x^{-1-\alpha/2}\,, \quad x \to \infty\,.$$

Ben Arous and Guionnet (2008), Belinschi et al. (2009)

- Assumption: X symmetric and regularly varying with index $\alpha \in (0,2)$.
- **Goal:** For $k \ge 1$, find the limit of

$$\mathbb{E}[\int x^k F_{\mathbf{R}}(dx)] = \frac{1}{p}\mathbb{E}[\mathrm{tr}(\mathbf{R}^k)]$$

Moments of LSD

One has

$$\mathbb{E}[\mathrm{tr}(\mathbf{R}^k)] = \sum_{i_1,\ldots,i_k=1}^{p} \underbrace{\sum_{t_1,\ldots,t_k=1}^{n} \mathbb{E}[Y_{i_1t_1}Y_{i_2t_1}\cdots Y_{i_kt_k}Y_{i_1t_k}]}_{:=F(i_1,\ldots,i_k)}.$$

Assumption: X symmetric $\Rightarrow Y_{ij}$ symmetric

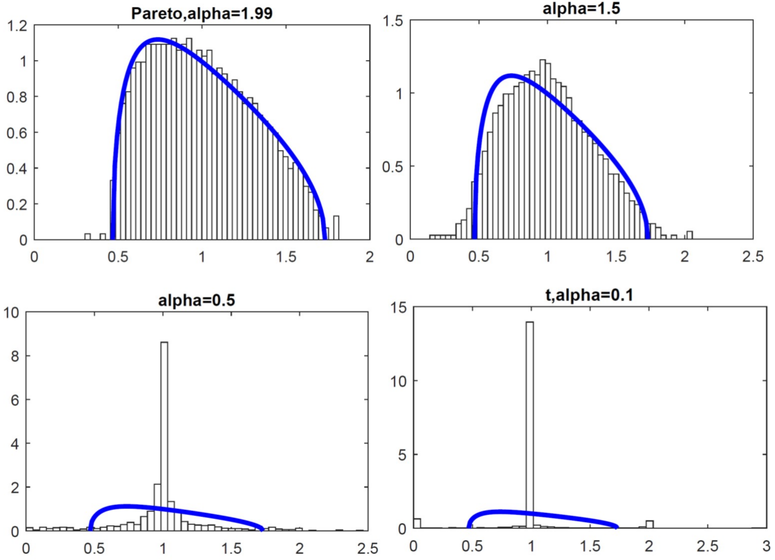
$$Y_{ij} = \frac{X_{ij}}{\sqrt{\sum_{t=1}^n X_{it}^2}}$$

Moments of LSD

$$\frac{1}{p}\mathbb{E}[\mathrm{tr}(\mathbf{R}^k)] \to \beta_k(\gamma) + \frac{2}{\alpha}\sum_{r=2}^{k-2}\gamma^{r-1}\sum_{q=0}^{r-2}(\Gamma(1-\alpha/2))^{-r+q+1}$$

$$\sum_{I\in\mathcal{C}_{r,k}^{(q)}}\sum_{s=1}^{t^\star(\widetilde{I})}\Big(\frac{\alpha/2}{\Gamma(1-\alpha/2)}\Big)^s\sum_{T\in\mathcal{C}_{s,|\widetilde{I}|}(\widetilde{I})}\left(\prod_{i=1}^{r-q}\frac{\Gamma(d_i(\widetilde{I},T))}{\Gamma(N_i(\widetilde{I}))}\right)$$

$$\prod_{(i,t)\in\Delta(\widetilde{I},T)}\Gamma\Big(\frac{m_{it}(\widetilde{I},T)-\alpha}{2}\Big)\,.$$



Pareto,alpha=1.99
alpha=1.5
alpha=0.5
t,alpha=0.1

Motivation

- **Random walk**

$$S_n = X_1 + \cdots + X_n\,, \qquad n \ge 1\,.$$

1. (X_i) are iid random variables with generic element X.
2. $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$.

- **Dimension** $p = p_n \to \infty$
- Consider iid copies $(S_n^{(i)})_{i\le p}$ of S_n and define the **point process**

$$N_n = \sum_{i=1}^{p} \delta_{d_p(S_n^{(i)}/\sqrt{n} - d_p)}\,.$$

We want to prove:

$$N_n = \sum_{i=1}^{p} \delta_{d_p(S_n^{(i)}/\sqrt{n}-d_p)} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,,$$

where N is a Poisson random measure with mean measure $\mu(x,\infty) = \mathrm{e}^{-x}$, $x \in \mathbb{R}$, and

$$d_p = \sqrt{2\log p} - \frac{\log\log p + \log 4\pi}{2(2\log p)^{1/2}}\,.$$

We want to prove:

$$N_n = \sum_{i=1}^{p} \delta_{d_p(S_n^{(i)}/\sqrt{n}-d_p)} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,,$$

where N is a Poisson random measure with mean measure $\mu(x,\infty) = \mathrm{e}^{-x}$, $x \in \mathbb{R}$, and

$$d_p = \sqrt{2\log p} - \frac{\log\log p + \log 4\pi}{2(2\log p)^{1/2}}\,.$$

- Note: d_p is the centering and normalizing sequence for the maximum of p iid standard normals.
- By Resnick (2007), this is equivalent to

$$p\,\mathbb{P}\big(d_p\,(S_n/\sqrt{n} - d_p) > x\big) \to \mathrm{e}^{-x}\,, \qquad x \in \mathbb{R}\,.$$

H., Mikosch, Yslas (2019+)

Assume that the sequence (p_n) satisfies the following conditions:

(C1) $p = O(n^{(s-2)/2})$ for $s > 2$ if $\mathbb{E}[|X|^s] < \infty$.

(C2) $p = \exp(o(n^{1/3}))$ if $\mathbb{E}[\exp(h\,|X|)] < \infty$ for some $h > 0$.

Then

$$p\,\mathbb{P}\big(d_p\,(S_n/\sqrt{n} - d_p) > x\big) \to \mathrm{e}^{-x}\,, \qquad x \in \mathbb{R}\,.$$

H., Mikosch, Yslas (2019+)

Assume that the sequence (p_n) satisfies the following conditions:

(C1) $p = O(n^{(s-2)/2})$ for $s > 2$ if $\mathbb{E}[|X|^s] < \infty$.

(C2) $p = \exp(o(n^{1/3}))$ if $\mathbb{E}[\exp(h\,|X|)] < \infty$ for some $h > 0$.

Then

$$p\,\mathbb{P}\big(d_p\,(S_n/\sqrt{n} - d_p) > x\big) \to \mathrm{e}^{-x}\,, \qquad x \in \mathbb{R}\,.$$

Precise large deviation bounds of the type

$$\sup_{0 \le y \le \gamma_n} \left| \frac{\mathbb{P}(S_n/\sqrt{n} > y)}{\overline{\Phi}(y)} - 1 \right| \to 0\,, \quad n \to \infty\,,$$

Under (C1): $\gamma_n = \sqrt{(s-2)\log n}$, Michel (1974)
Under (C2): $\gamma_n = o(n^{1/6})$, Petrov (1972)

- **Data matrix** $\boldsymbol{X} = \boldsymbol{X}_n$: $p \times n$ matrix with iid entries with generic element X.

$$\boldsymbol{X} = (X_{it})_{i=1,\ldots,p;t=1,\ldots,n}$$

- **Sample covariance matrix**

$$\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}'$$

Dependent random walks

$$S_{ij} = \sum_{t=1}^{n} X_{it} X_{jt} \,, \quad i < j \,.$$

- **Off-diagonal point process:**

$$N_n^S = \sum_{1 \le i < j \le p} \delta_{\widetilde{d}_p (S_{ij}/\sqrt{n} - \widetilde{d}_p)} \,,$$

where $\widetilde{d}_p = d_{p(p-1)/2}$.

Off-diagonal point process

$$N_n^S = \sum_{1\le i<j\le p} \delta_{\widetilde{d}_p(S_{ij}/\sqrt{n}-\widetilde{d}_p)}$$

Theorem: H., Mikosch, Yslas (2019+)

Assume that the sequence (p_n) satisfies:

- $p = O(n^{(s-2)/4})$ for $s > 2$ if $\mathbb{E}[|X|^s] < \infty$.
- $p = \exp(o(n^{1/3}))$ if $\mathbb{E}[\exp(h\,|X_{11}X_{12}|)] < \infty$ for some $h > 0$.

Then

$$N_n^S \xrightarrow{\mathrm{d}} N\,.$$

Remark: Entries of $\mathbf{X}$ do not have to be identically distributed.

- Note that
$$N=\sum_{i=1}^{\infty}\delta_{-\log\Gamma_i}\,,$$
where $\Gamma_i=E_1+\cdots+E_i$, $i\geq 1$, and (E_i) is iid standard exponential.

- Note that
$$N = \sum_{i=1}^{\infty} \delta_{-\log \Gamma_i}\,,$$
where $\Gamma_i = E_1 + \cdots + E_i$, $i \ge 1$, and (E_i) is iid standard exponential.
- For fixed k,
$$\widetilde{d}_p\big(S_{(i)}/\sqrt{n} - \widetilde{d}_p\big)_{i=1,\ldots,k} \xrightarrow{\mathrm{d}} (-\log \Gamma_i)_{i=1,\ldots,k}\,.$$

- In particular, Jiang (2004)

$$\lim_{n\to\infty} \mathbb{P}\Big(\widetilde{d}_p\big(S_{(1)}/\sqrt{n} - \widetilde{d}_p\big) \le x\Big) = \exp(-\mathrm{e}^{-x})\,.$$

- Fang Han's talk on Tuesday, Songxi Chen's talk on Wednesday

Extension to sample correlation matrices

Sample correlation matrix

$$\mathbf{R} = (\mathrm{diag}(\boldsymbol{S}))^{-1/2}\boldsymbol{S}(\mathrm{diag}(\boldsymbol{S}))^{-1/2}$$

Extension to sample correlation matrices

Sample correlation matrix

$$\mathbf{R} = (\mathrm{diag}(\boldsymbol{S}))^{-1/2}\boldsymbol{S}(\mathrm{diag}(\boldsymbol{S}))^{-1/2}$$

Theorem: H., Mikosch, Yslas (2019+)

Assume that the sequence (p_n) satisfies:

- $p = O(n^{(s-2)/4})$ for $s > 2$ if $\mathbb{E}[|X|^s] < \infty$.
- $p = \exp(o(n^{1/3}))$ if $\mathbb{E}[\exp(h\,|X_{11}X_{12}|)] < \infty$ for some $h > 0$.

Then

$$N_n^R = \sum_{1 \le i < j \le p} \delta_{\widetilde{d}_p(\sqrt{n}R_{ij} - \widetilde{d}_p)} \xrightarrow{\mathrm{d}} N\,.$$

Thank you!

(Z_{it}): iid field of regularly varying random variables.

- **Stochastic volatility model**:

$$\boldsymbol{X} = \big(Z_{it}\,\sigma_{it}^{(n)}\big)_{p\times n}$$

Heavy Tails and Dependence

(Z_{it}): iid field of regularly varying random variables.

- **Stochastic volatility model**:

$$\boldsymbol{X} = \big(Z_{it}\,\sigma_{it}^{(n)}\big)_{p\times n}$$

- **Generate deterministic covariance structure $\mathbf{A}$**:

$$\boldsymbol{X} = \mathbf{A}^{1/2}\mathbf{Z}$$

Davis et al. (2014)

(Z_{it}): iid field of regularly varying random variables.

- **Dependence among rows and columns**:

$$X_{it} = \sum_{l=0}^{\infty} \sum_{k=0}^{\infty} h_{kl} Z_{i-k,t-l}$$

with some constants h_{kl}. Davis et al. (2016)

Heavy Tails and Dependence

(Z_{it}): iid field of regularly varying random variables.

- **Dependence among rows and columns**:

$$X_{it} = \sum_{l=0}^{\infty} \sum_{k=0}^{\infty} h_{kl} Z_{i-k,t-l}$$

  with some constants h_{kl}. Davis et al. (2016)

- **Relation to iid case:**

$$\boldsymbol{X}\boldsymbol{X}' = \sum_{l_1,l_2=0}^{\infty} \sum_{k_1,k_2=0}^{\infty} h_{k_1 l_1} h_{k_2 l_2} \mathbf{Z}(k_1, l_1) \mathbf{Z}'(k_2, l_2) \,,$$

  where

$$\mathbf{Z}(k,l) = (Z_{i-k,t-l})_{i=1,\ldots,p;t=1,\ldots,n} \,, \quad l, k \in \mathbb{Z} \,.$$

Heavy Tails and Dependence

(Z_{it}): iid field of regularly varying random variables.

- **Dependence among rows and columns**:

$$X_{it} = \sum_{l=0}^{\infty} \sum_{k=0}^{\infty} h_{kl} Z_{i-k,t-l}$$

with some constants h_{kl}. Davis et al. (2016)

- **Relation to iid case:**

$$\boldsymbol{X}\boldsymbol{X}' = \sum_{l_1,l_2=0}^{\infty} \sum_{k_1,k_2=0}^{\infty} h_{k_1 l_1} h_{k_2 l_2} \mathbf{Z}(k_1, l_1) \mathbf{Z}'(k_2, l_2)\,,$$

where

$$\mathbf{Z}(k,l) = (Z_{i-k,t-l})_{i=1,\dots,p;t=1,\dots,n}\,, \quad l,k \in \mathbb{Z}\,.$$

- **Location of squares:**

$$\boldsymbol{M}_{ij} = \sum_{l \in \mathbb{Z}} h_{il} h_{jl}, \qquad i,j \in \mathbb{Z}\,.$$

Autocovariance Matrices

- For $s \geq 0$,

$$\boldsymbol{X}_n(s) = (X_{i,t+s})_{i=1,\ldots,p;\, t=1,\ldots,n}\,, \quad n \geq 1\,.$$

  Then $\boldsymbol{X}_n = \boldsymbol{X}_n(0)$.

- **Autocovariance matrix** for lag s

$$\boldsymbol{X}_n(0)\boldsymbol{X}_n(s)'$$

- Limit theory for singular values of such matrices.

Autocovariance Matrices

- **Autocovariance matrix** for lag s

$$\mathbf{C}_n(s) = \begin{cases} \boldsymbol{X}_n(0)\boldsymbol{X}_n(s)', & \text{if } \alpha < 2(1+\beta), \\ \boldsymbol{X}_n(0)\boldsymbol{X}_n(s)' - \mathbb{E}[\boldsymbol{X}_n(0)\boldsymbol{X}_n(s)'], & \text{if } \alpha > 2(1+\beta), \end{cases}$$

- Consider

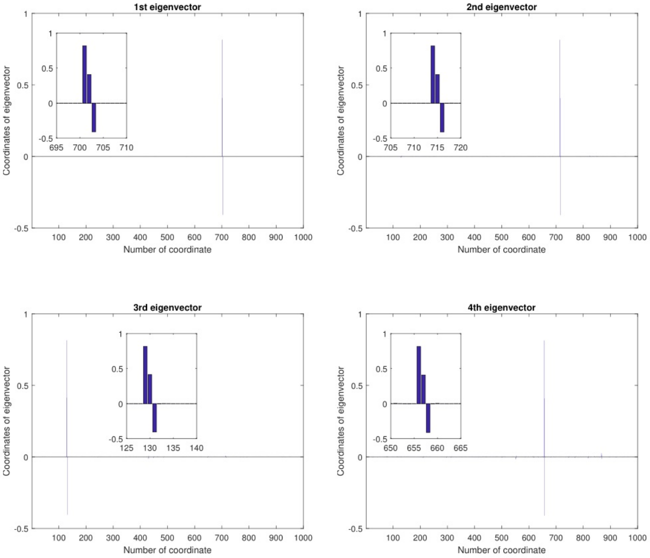
$$\mathbf{P}_n(s_1, s_2) = \sum_{s=s_1}^{s_2} \mathbf{C}_n(s)\mathbf{C}_n(s)' \quad \text{for fixed } 0 \le s_1 \le s_2.$$

Autocovariance Matrices

- $$(\boldsymbol{M}(s))_{ij} = \sum_{l\in\mathbb{Z}} h_{i,l}h_{j,l+s}, \qquad i,j \in \mathbb{Z}.$$

  For $0 \le s_1 \le s_2 < \infty$, we define the positive semi-definite matrix
  $$\mathbf{K}(s_1,s_2) = \sum_{s=s_1}^{s_2} \boldsymbol{M}(s)\boldsymbol{M}(s)'$$
- Eigenvector approximation
  $$\|\boldsymbol{y}_i(s_1,s_2) - \mathbf{u}_{b(i)}^{a(i)}(s_1,s_2)\|_{\ell_2} \xrightarrow{\mathbb{P}} 0\,, \qquad n\to\infty\,.$$

Autocovariance eigenvectors



Autocovariance eigenvectors