Statistical Foundations of Data Science

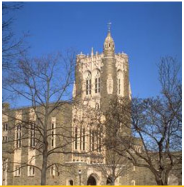
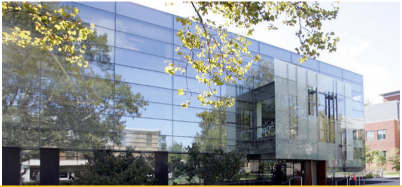
Jianqing Fan

Princeton University

https://fan.princeton.edu

ZOOM ID Lectures: **970 4936 8998**   Office Hours: **996 4030 7631**

Annotated Lecture Notes: web view


8. Covariance Regularization and Graphical Models

8.1. Matrix norms (§9.1)

8.2. Sparse Covariance Estimation (§9.2)

8.3. Robust Covariance Inputs (§9.3)

8.4 Sparse Precision Matrix and Graphical Models (§9.4)

8.1. Matrix Norms

Norms of Matrices

Definition: A norm of an $n \times m$ matrix satisfies

1. $\|\mathbf{A}\| \geq 0$ and $\|\mathbf{A}\| = 0$ iff $\mathbf{A} = 0$;
2. $\|\alpha \mathbf{A}\| = |\alpha| \|\mathbf{A}\|$;
3. $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$;
4. $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|$.

Induced norm: $\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|$.

(p,q)-norm: $\|\mathbf{A}\|_{p,q} = \max_{\|\mathbf{x}\|_p=1} \|\mathbf{Ax}\|_q$. Hence,

$$\|\mathbf{Ax}\|_q \leq \|\mathbf{A}\|_{p,q} \|\mathbf{x}\|_p.$$

L_p-Norms of Matrices

■ $\|\mathbf{A}\|_p = \|\mathbf{A}\|_{p,p} = \max_{\|\mathbf{x}\|_p=1} \|\mathbf{A}\mathbf{x}\|_p$.

1. Operator norm: $\|\mathbf{A}\|_2 = \lambda(\mathbf{A}^T\mathbf{A})^{1/2}$.
2. L_1-norm: $\|\mathbf{A}\|_1 = \max_j \sum_{i=1}^m |a_{ij}|$, max L_1-norm of columns.
3. L_∞-norm: $\|\mathbf{A}\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$, max L_1-norm of rows.

Frobenius norm: $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \sum_{i,j} a_{ij}^2 = |\lambda_1|^2 + \cdots + |\lambda_m|^2$.

Nuclear norm: $\|\mathbf{A}\|_* = |\lambda_1| + \cdots + |\lambda_m|$.

Inequalities

1. $n^{-1/2}\|\mathbf{A}\|_\infty \le \|\mathbf{A}\|_2 \le m^{1/2}\|\mathbf{A}\|_\infty$
   $m^{-1/2}\|\mathbf{A}\|_1 \le \|\mathbf{A}\|_2 \le n^{1/2}\|\mathbf{A}\|_1$

2. $\|\mathbf{A}\|_2^2 \le \|\mathbf{A}\|_\infty \|\mathbf{A}\|_1$, $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_1$ **if A symmetric**.

3. $\|A\|_{max} \le \|\mathbf{A}\|_2 \le (mn)^{1/2}\|A\|_{\max}$, ★$\|\mathbf{A}\|_{\max} = \max_{ij}|a_{ij}|$

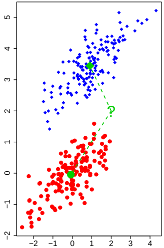
4. $\|\mathbf{A}\| \le \|\mathbf{A}\|_F \le r\|\mathbf{A}\|$, $r = \mathrm{rank}(\mathbf{A})$.

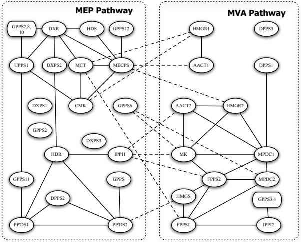
8.2 Sparse Covariance Estimation

Needs of Covariance

Finance: ★Risk estimation; ★Portfolio choices; ★Factor models; ★PCA Reg;

Machine Learning: ★Classification; ★network; ★topic modeling; ★matrix completion





Graphical Modeling: Conditional dependence modeling

Staistical Inferences: ★FDR controls; ★General LS; ★Regression

$$E(Y - \mathbf{X}^T\beta)^2 = (-\beta^T, 1)\mathbf{\Sigma}^*(-\beta^T, 1)^T, \text{ where } \mathbf{\Sigma}^* = \operatorname{cov}((\mathbf{X}^T, Y)^T).$$

Classical Multivariate Analysis

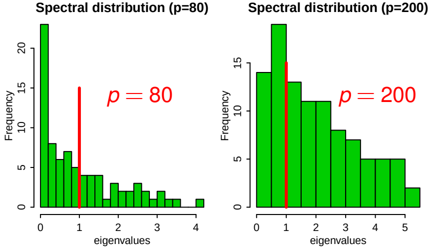
$p : 3 \sim 8$, $n = 30$ — 100

Asymptotic framework: $n \to \infty$, but p **fixed**.

- inappropriate for many contemporary applications.
- more appropriate: $p \to \infty$ and $n \to \infty$ and study impact of p

★High-dim: $p = 3K$ gives 4.5m parameters

★unexpected behavior and degeneracy



■Spectral dist for sample cov matrix from $N(0, \mathbf{I}_p)$ with $n = 100$.
■Theoretical: point mass at 1.

Sparse covariance estimation

Gaussian likelihood: $Q(\boldsymbol{\Sigma}) = -\log|\boldsymbol{\Sigma}^{-1}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})$,

like = $|\boldsymbol{\Sigma}|^{-n/2}\exp(-\frac{1}{2}\sum_{i=1}^{n} \quad (\mathbf{x}_i - \mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mu))$

Penalized QMLE: min $Q(\boldsymbol{\Sigma}) + \sum_{i\neq j} p_\lambda(|\sigma_{ij}|)$.

★hard to compute

Thresholding: $\widehat{\boldsymbol{\Sigma}}_\lambda = \left(s_{i,j}\mathbb{I}(|s_{i,j}| \geq \lambda)\right)$,

★$\lambda =$ thresholding parameter.

It solves $\sum_{i,j}(\sigma_{ij} - s_{ij})^2 + \sum_{i\neq j} p_\lambda(|\sigma_{ij}|)$.

(Bickel and Levina, 08a)

■Therehsolding reduces variance

■sparsity controls biases

Banding: For banded structure *(Bickel and Levina, 08b)*,

$$\widehat{\boldsymbol{\Sigma}}_k^B = (s_{ij} I(|i-j| \geq k))$$

Variations of threholding estimator

1. Generalized threholding (*Rothman, et al, 09*):

$$\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}} = (\tau_\lambda(\widehat{\sigma}_{ij})) \qquad \bullet\tau_\lambda(\cdot) \text{ by } (\textit{Antoniadis and Fan, 01}).$$

   a) $|\tau_\lambda(z)| \le a|y|$ for all $z,\ y$ that satisfy $|z-y| \le \lambda/2$; $\Longrightarrow \tau_\lambda(z) = 0$ for $|z| \le \lambda/2$.

   b) $|\tau_\lambda(z) - z| \le \lambda$, for all $z \in \mathbb{R}$.

2. Adaptive thresholding (*Cai and Liu, 11*): $\widehat{\boldsymbol{\Sigma}}_\lambda = \left(\widehat{\sigma}_{i,j}\mathbb{I}\left(\frac{|\widehat{\sigma}_{i,j}|}{\mathrm{SE}(\widehat{\sigma}_{i,j})} \ge \lambda\right)\right)$.

3. Entry dependent thresholding (*Fan, Liao, Mincheva, 11*)

$$\widehat{\boldsymbol{\Sigma}}^{\tau}_{\lambda} = \left(\tau_{\lambda_{ij}}(\widehat{\sigma}_{ij})\right) \equiv \left(\widehat{\sigma}^{\tau}_{ij}\right), \qquad \lambda_{ij} = \lambda\sqrt{\widehat{\sigma}_{i,i}\widehat{\sigma}_{j,j}\frac{\log p}{n}}.$$

★ $\equiv$ thresholding at correlation

★diag when $\lambda = 1/\sqrt{(\log p)/n}$

Class of Sparse Covariance Matrices

Controll operator-norm: $\|\widehat{\mathbf{\Sigma}}^{\tau}_{\lambda} - \mathbf{\Sigma}\|_2 \le \max_i \sum_{j=1}^{p} |\widehat{\sigma}^{\tau}_{ij} - \sigma_{ij}|$.

★error depends on sparsity measure $m_{p,0} = \max_{i \le p} \sum_{j=1}^{p} \mathbb{I}\{\sigma_{ij} \neq 0\}$ (L_0-norm).

Generalized measure of sparsity: $m_{p,q} = \max_{i \le p} \sum_{j=1}^{p} |\sigma_{ij}|^q, \qquad q < 1.$

Parameter space: $\{\mathbf{\Sigma} \succeq 0 : \sigma_{ii} \le C, \sum_{j=1}^{p} |\sigma_{ij}|^q \le m_p\}$ is generalized to

$$\mathcal{C}_q(m_p) = \left\{ \mathbf{\Sigma} : \max_i \sum_j (\sigma_{ii}\sigma_{jj})^{(1-q)/2} |\sigma_{ij}|^q \le C^{1-q} m_p \right\},$$

since $\|\mathbf{\Sigma}\|_2 \le \max_i \sum_j |\sigma_{ij}| \le \max_i \sum_j (\sigma_{ii}\sigma_{jj})^{(1-q)/2} |\sigma_{ij}|^q \le C^{1-q} m_p$.

Asymptotic Property (I)

Theorem 9.1. If $\sup_{\boldsymbol{\Sigma}\in\mathcal{C}_q(m_p)} P\left(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} > C_0\sqrt{(\log p)/n}\right) \le \varepsilon_{n,p}$,

$\log p = o(n)$ and $\min_{i\le p}\sigma_{ii} = \gamma > 0$, then

$$\sup_{\boldsymbol{\Sigma}\in\mathcal{C}_q(m_p)} P\left\{\|\widehat{\boldsymbol{\Sigma}}_\lambda^\tau - \boldsymbol{\Sigma}\|_2 > C_1 m_p\left(\frac{\log p}{n}\right)^{(1-q)/2}\right\} \le 3\varepsilon_{n,p}.$$

For Frobenius norm, if $\max_{i\le p}\sigma_{ii} \le C_2$

$$p^{-1}\|\widehat{\boldsymbol{\Sigma}}_\lambda^\tau - \boldsymbol{\Sigma}\|_F^2 = O_P\left(m_p\left(\frac{\log p}{n}\right)^{1-q/2}\right).$$

In addition, if $\|\boldsymbol{\Sigma}^{-1}\|$ is bounded from below, then

$$\left\|\left(\widehat{\boldsymbol{\Sigma}}_\lambda^\tau\right)^{-1} - \boldsymbol{\Sigma}^{-1}\right\| = O_P\left(m_p\left(\frac{\log p}{n}\right)^{(1-q)/2}\right).$$

Remarks

1. Determistic result from input accuracy + sparsity structure. Rates are stated for different norms and optimal.

2. When $q = 0$, the rates are $m_p \left(\frac{\log p}{n}\right)^{1/2}$, as expected.

3. For subGaussian data, $\|\mathbf{S} - \mathbf{\Sigma}\|_{\max} = O_P\left(\sqrt{\frac{\log p}{n}}\right)$.

   SubGaussianity: $\kappa = \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{v}^T \mathbf{X}_i\|_{\psi_2} < \infty$, where $\|X\|_{\psi_2} = \inf\{s > 0 : \quad E\exp(X^2/s^2) \leq 2\}$ (Orlicz norm).

4. Rate $\sqrt{\frac{\log p}{n}}$ in the theorem can be replaced by any $a_n \to 0$.

Proof of Theorem 9.1

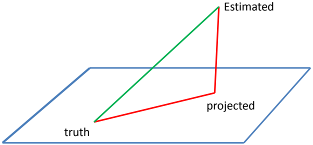
Let events $E_1 \equiv \{|\widehat{\sigma}_{ij} - \sigma_{ij}| \le \lambda_{ij}/2, \forall i, j\}$ and $E_2 = \{\widehat{\sigma}_{ii}\widehat{\sigma}_{jj} \le 2\sigma_{ii}\sigma_{jj}, \forall i, j\}$.

1. On E_1, we have $\left|\tau_{\lambda_{ij}}(\widehat{\sigma}_{ij}) - \sigma_{ij}\right| \le \left|\tau_{\lambda_{ij}}(\widehat{\sigma}_{ij}) - \widehat{\sigma}_{ij}\right| + \left|\widehat{\sigma}_{ij} - \sigma_{ij}\right| \le 1.5\lambda_{ij}$,
2. Using property a) of $\tau_\lambda(\cdot)$, we have

$$\begin{aligned} \left|\tau_{\lambda_{ij}}(\widehat{\sigma}_{ij}) - \sigma_{ij}\right| &\le 1.5\lambda_{ij} 1\{|\sigma_{ij}| \ge \lambda_{ij}\} + (1+a)\sigma_{ij} 1\{|\sigma_{ij}| < \lambda_{ij}\} \\ &\le 1.5|\sigma_{ij}|^q \lambda_{ij}^{1-q} + (1+a)|\sigma_{ij}|^q \lambda_{ij}^{1-q} = (2.5+a)|\sigma_{ij}|^q \lambda_{ij}^{1-q} \end{aligned}$$

3. Hence, $\sum_{j=1}^p \left|\tau_{\lambda_{ij}}(\widehat{\sigma}_{ij}) - \sigma_{ij}\right| \le (2.5+a)\sum_{j=1}^p \lambda_{ij}^{1-q}|\sigma_{ij}|^q$
$\le (2.5+a)(2\lambda)^{1-q}\left(\frac{\log p}{n}\right)^{(1-q)/2} \sum_{j=1}^p (\sigma_{ii}\sigma_{jj})^{(1-q)/2}|\sigma_{ij}|^q$ on E_2.
4. On $\mathcal{C}_q(m_p)$, $\|\widehat{\mathbf{\Sigma}}_\lambda^\tau - \widehat{\mathbf{\Sigma}}\|_2 \le \max_i \sum_{j=1}^p |\widehat{\sigma}_{ij}^\tau - \sigma_{ij}| \le C_1 m_p \left(\frac{\log p}{n}\right)^{(1-q)/2}$,
5. The Frobenius norm follows from the same calculation.

Projection of symmetric matrices



Problem: $\widehat{\boldsymbol{\Sigma}}_\lambda$ is not necessarily positive definite.

★ **Method 1**: Set $\widehat{\boldsymbol{\Sigma}}_\lambda^+ = \boldsymbol{\Gamma}^T \mathrm{diag}(\lambda_1^+, \cdots, \lambda_p^+)\boldsymbol{\Gamma}$.

★ **Method 2**: (still a corr matrix) $\widehat{\boldsymbol{\Sigma}}_\lambda^+ = (\widehat{\boldsymbol{\Sigma}}_\lambda + \lambda_{\min}^- I_p)/(1+\lambda_{\min}^-)$.

■ Both projections do not alter eigenvectors.

★ **Method 3**: Nearest positive definite projection:

$$\|\mathbf{A}-\mathbf{R}\|_F^2, \quad \text{s.t.} \quad \lambda_{\min}(\mathbf{R}) \ge \delta, \mathrm{diag}(\mathbf{R}) = \mathbf{I}_p.$$

for a given $\delta \ge 0$.

★ `nearPD` in R-package NearPD computes this.

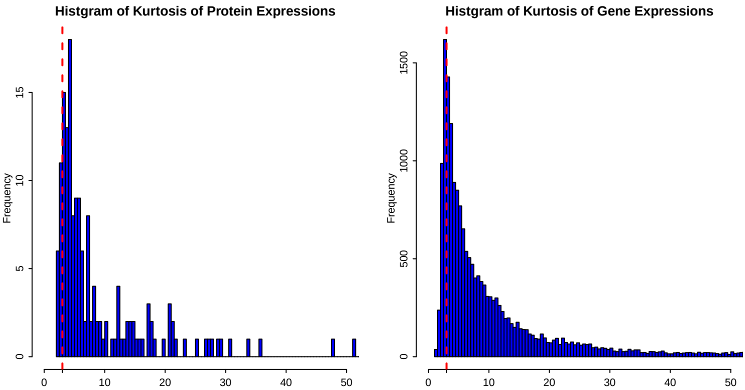
8.3 Robust Covariance Inputs

Heavy-tailed distributions

- ubiquitous in modern statistics and machine learning
  - ★ financial returns; macroeconomics time series
  - ★ high-throughput data: microarrays, proteomics, fMRI
  - ★ arising easily in high-dimensional data
- at odd with sub-Gaussian or sub-exponential assumptions

Example: Protein and Gene Expressions

■ NCI-60: 60 human cancer cell lines (Shankavaram et al., 2007)

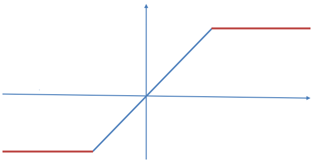


Protein: 49/162 **Gene**: 6542/17924 heavier than t_5!

Principle of Robustification (I): Truncation

Data: $X_i \sim \text{IID}(\mu, \sigma^2)$.

Truncation: Let $\widetilde{X}_i = \text{sgn}(X_i)\min(|X_i|, \tau)$.



Exponential concentration: When $\tau \asymp \sigma\sqrt{n}$, (Fan, Wang, Zhu, 20)

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i - \mu\right| \geq t\frac{\sigma}{\sqrt{n}}\right) \leq 2\exp(-ct^2), \quad \text{univ const } c$$

$$\preceq 1/t^2, \quad \text{for } \textbf{sample mean}$$

- Truncated mean behaves like **Gaussian**, whereas ave like **Cauchy**.
- Fundamental to high-dim. estimation

Robust Covariance Inputs by Truncation

Data: $\mathbf{X}_i \sim IID(0, \mathbf{\Sigma})$, $\quad p$-dim. $\qquad \sigma_{ij} = E(X_i X_j)$

Robust Covariance: ave of truncated data: $\widetilde{\mathbf{\Sigma}} = \left(n^{-1} \sum_{k=1}^{n} \widetilde{\mathbf{x}}_{k,ij}\right)$

Elementwise truncation: $\widetilde{x}_{k,ij} = \mathrm{sgn}(x_{k,i} x_{k,j}) max(|x_{k,i} x_{k,j}|, \tau)$ at τ

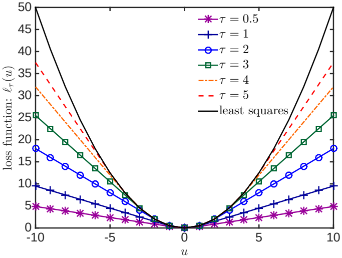
■ Assuming bounded fourth moments, we have $\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} = O\left(\sqrt{\frac{\log \mathbf{p}}{\mathbf{n}}}\right)$

Robust U-covariance: Note $\mathbf{\Sigma} = \frac{1}{2} E\|\mathbf{X}_i - \mathbf{X}_j\|^2 \frac{(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^T}{\|\mathbf{X}_i - \mathbf{X}_j\|^2}$

$$\widehat{\mathbf{\Sigma}}_U = \frac{1}{2\binom{n}{2}} \sum_{i \neq j} \min(\|\mathbf{X}_i - \mathbf{X}_j\|^2, \tau) \frac{(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^T}{\|\mathbf{X}_i - \mathbf{X}_j\|^2}$$

Principle of High-dim Robustification II

■ Adaptive Huber loss:



g 17)

$$\rho_\tau(x) = \begin{cases} x^2, & \text{if } |x| \le \tau \\ \tau(2|x| - \tau), & \text{if } |x| > \tau \end{cases}.$$

$$\widehat{\mu}_\tau = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau(Y_i - \mu),$$

★ More convenient for regression

★ Same concentration property holds (Fan, Li, Wang 17)

For $\tau = \sqrt{nc}/t$ with $c \ge \mathrm{SD}(Y)$, (Fan, Li, Wang 17)

$$P(|\widehat{\mu}_\tau - \mu| \ge t\frac{c}{\sqrt{n}}) \quad \le \quad 2\exp(-\mathbf{t^2/16}), \quad \forall t \le \sqrt{n/8}$$

Robust Covariance Inputs (II)

Elementwise estimator: $\widehat{\mathbf{\Sigma}}_E = \left(\widehat{\mathrm{E}(X_i X_j)}^a - \widehat{\mathrm{E}\, X_i}^a \widehat{\mathrm{E}\, X_j}^a \right)$

★ $a = T$ means “truncation” ★ $a = H$ refers to Huber estimator

If 4^{th} moment uniformly bounded, $\|\widehat{\mathbf{\Sigma}}_E - \mathbf{\Sigma}\|_{\max} = O_P\left(\sqrt{\frac{\log p}{n}}\right)$.

Other methods: ★shrinkage ★Rank-correlation

8.4 Sparse Precision Matrix and Graphical Models

Graussian graphical models

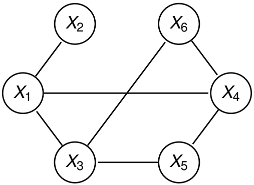
Model: Let $\mathbf{X} \sim N(\mu, \boldsymbol{\Sigma})$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}$$

Then, $\boldsymbol{\Omega}_{11} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}\boldsymbol{\Sigma}_{21}$ and

$$(\mathbf{X}_1 | \mathbf{X}_2) \sim N\big(\mu_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}(\mathbf{X}_2 - \mu_2), \boldsymbol{\Omega}_{11}\big)$$

$\omega_{ij} = 0 \Longleftrightarrow X_i$ and X_j are conditionally indep, given rest variables



$$\Omega = \begin{bmatrix} 1 & 0.2 & 0.2 & 0.2 & 0 & 0 \\ 0.2 & 1 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 1 & 0 & 0.2 & 0.2 \\ 0.2 & 0 & 0 & 1 & 0.2 & 0.2 \\ 0 & 0 & 0.2 & 0.2 & 1 & 0 \\ 0 & 0 & 0.2 & 0.2 & 0 & 1 \end{bmatrix}$$

■ Sparsity patten of $\boldsymbol{\Omega}$ is depicted by graph (no magnitude)

Penalized MLE and Least-squares

PMLE: $\text{argmin}_{\boldsymbol{\Omega} \succ 0}\{-\log|\boldsymbol{\Omega}| + \text{tr}(\boldsymbol{\Omega}\mathbf{S}) + \sum_{i \neq j} p_{\lambda_{ij}}(|\omega_{ij}|)\}$.

Proposition 9.4. Let α_j^* and β_j^* be the solution to least-squares

$$\min E(X_j - \alpha_j - \beta_j^T \mathbf{X}_{-j})^2 \qquad \tau_j^* = \min E(X_j - \alpha_j - \beta_j^T \mathbf{X}_{-j})^2.$$

Then, j^{th} column of $\boldsymbol{\Omega}^*$ is given by

col-by-col solution

$$\omega_{jj}^* = 1/\tau_j^*, \qquad \omega_j^* = \beta_j^*/\tau_j^*.$$

PLS: $\sum_{i=1}^n (X_{ij} - \alpha - \beta^T \mathbf{X}_{i,-j})^2 + \sum_{k=1}^{p-1} p_\lambda(|\beta_k|)$. *(Meinshausen & Buehlmann, 06)*

sqrt lasso: $\{\sum_{i=1}^n (X_{ij} - \alpha - \beta^T \mathbf{X}_{i,-j})^2\}^{1/2} + \lambda\|\beta\|_1$ *(Belloni, Chernozhukov, Wang, 11)*

★scale-free ★see network figure.

CLIME

Constrained L_1-minimization for Inverse Matrix Estimation:

$$\min \sum_{j=1}^{p} \|\omega_j\|_1, \qquad \text{subject to } \|\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \mathbf{I}_p\|_{\max} \le \lambda_n, \qquad \textit{(Cai, Liu and Luo, 11)}$$

★Danzig selector (*Candes and Tao, 07)*

- Solving col-by-col: $\min \|\omega_j\|_1, s.t. \|\widetilde{\boldsymbol{\Sigma}}\omega_j - \mathbf{e}_j\|_\infty \le \lambda_n$. ★LP
- Solution is not necessary symmetric. Take the symmetric one with smaller magnitude:

$$\widehat{\Omega}_s = (\widehat{\omega}_{ij}^1 I(|\widehat{\omega}_{ij}^1| \le |\widehat{\omega}_{ji}^1|) + \widehat{\omega}_{ji}^1 I(|\widehat{\omega}_{ji}^1| < |\widehat{\omega}_{ij}|).$$

Statistical properties

Parameter space: $\mathcal{C}_q^*(m_p) = \{\boldsymbol{\Omega} \succ 0 : \|\boldsymbol{\Omega}\|_1 \le D_n, \quad \sum_{j=1}^p |\omega_{ij}|^q \le m_p\}$.

Theorem 9.6. If $\lambda_n \ge \|\boldsymbol{\Omega}^*\|_1 \|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\max}$, then uniformly in $\boldsymbol{\Omega}^* \in \mathcal{C}_q^*(m_p)$

$$\begin{aligned}
\|\widehat{\boldsymbol{\Omega}}_s - \boldsymbol{\Omega}^*\|_{\max} &\le 4\|\boldsymbol{\Omega}^*\|_1 \lambda_n \\
\|\widehat{\boldsymbol{\Omega}}_s - \boldsymbol{\Omega}^*\|_2 &\le 12 m_p (4\|\boldsymbol{\Omega}^*\|_1 \lambda_n)^{1-q}, \\
\frac{1}{p}\|\widehat{\boldsymbol{\Omega}}_s - \boldsymbol{\Omega}^*\|_F^2 &\le 12 m_p (4\|\boldsymbol{\Omega}^*\|_1 \lambda_n)^{2-q}.
\end{aligned}$$

If $\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = O_P(\sqrt{(\log p)/n})$, then we can take $\lambda_n = C_0 D_n \sqrt{(\log p)/n}$ and have explicit rates.

- Use matrix structure + elementwise convergence

Outline of Proof*

1. Verify $\mathbf{\Omega}^*$ satisfies the constraint $\Longrightarrow \|\widehat{\mathbf{\Omega}}_s\|_1 \leq \|\widehat{\mathbf{\Omega}}\|_1 \leq \|\mathbf{\Omega}^*\|_1$
2. 1^{st} follows from $\|\widehat{\mathbf{\Omega}}_s - \mathbf{\Omega}^*\|_{\max} \leq \|\mathbf{\Omega}^*\|_1 \|\mathbf{\Sigma}^*(\widehat{\mathbf{\Omega}}_s - \mathbf{\Omega}^*)\|_{\max}$ and

$$\|\mathbf{\Sigma}^*(\widehat{\mathbf{\Omega}}_s - \mathbf{\Omega}^*)\|_{\max} \leq \|\widetilde{\mathbf{\Sigma}}(\widehat{\mathbf{\Omega}}_s - \mathbf{\Omega}^*)\|_{\max} + \|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)(\widehat{\mathbf{\Omega}}_s - \mathbf{\Omega}^*)\|_{\max}.$$

   1^{st}-term bounded by $2\lambda_n$ by inserting $\mathbf{I}_p$ and 2^{nd}-term by $2\|\mathbf{\Omega}^*\|_1 \|\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\max} \leq 2\lambda_n$.

3. To prove 2^{nd} result, let $\widehat{\mathbf{\Omega}}_1 = (\widehat{\omega}^s_{ij} I(|\widehat{\omega}^s_{ij}| > 2a_n))$ and $\widehat{\mathbf{\Omega}}_2 = \widehat{\mathbf{\Omega}}_s - \widehat{\mathbf{\Omega}}_1$, where $a_n = \|\widehat{\mathbf{\Omega}}_s - \mathbf{\Omega}^*\|_{\max}$. Then,

$$\|\widehat{\omega}^1_j\|_1 + \|\widehat{\omega}^2_j\|_1 = \|\widehat{\omega}^s_j\|_1 \leq \|\widehat{\omega}_j\|_1 \leq \|\omega^*_j\|_1.$$

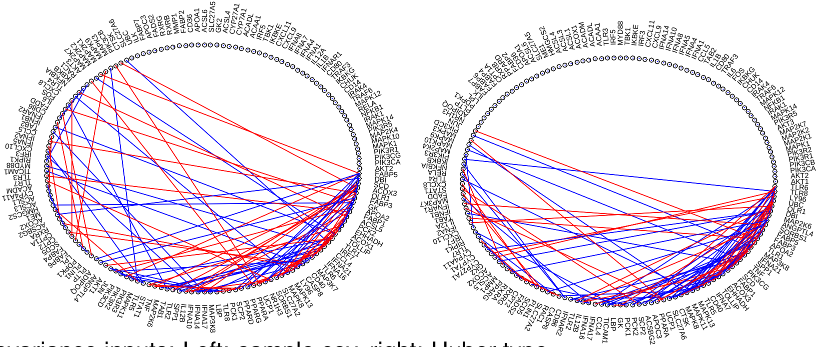
   Also, $\|\widehat{\omega}^1_j\|_1 \geq \|\omega^*_j\|_1 - \|\widehat{\omega}^1_j - \omega^*_j\|_1 \Longrightarrow \|\widehat{\omega}^2_j\|_1 \leq \|\widehat{\omega}^1_j - \omega^*_j\|_1$. **Conclusion**: $\|\widehat{\mathbf{\Omega}}_s - \mathbf{\Omega}^*\|_1 \leq 2\|\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}^*\|_1$.

4. Remaining follows from sparsity + thresholding

An illustration

Data: 95 genes from TLR pathway (related to cardiovascular disease) & 68 genes from PPAR pathway (unrelated to the disease), $n = 48$.

Parameter: λ chosen to have 100 connections



★Covariance inputs: Left: sample cov, right: Huber type

★blue: within pathway connections; red between connects.

★RACLIME: within = 60, between = 40; ACLIME: within = 55, between = 45