Sparse Matrix Graphical Models

(2012) @ JASA

by Chenlei Leng & Cheng Yong Tang

Presented by Javier Zapata

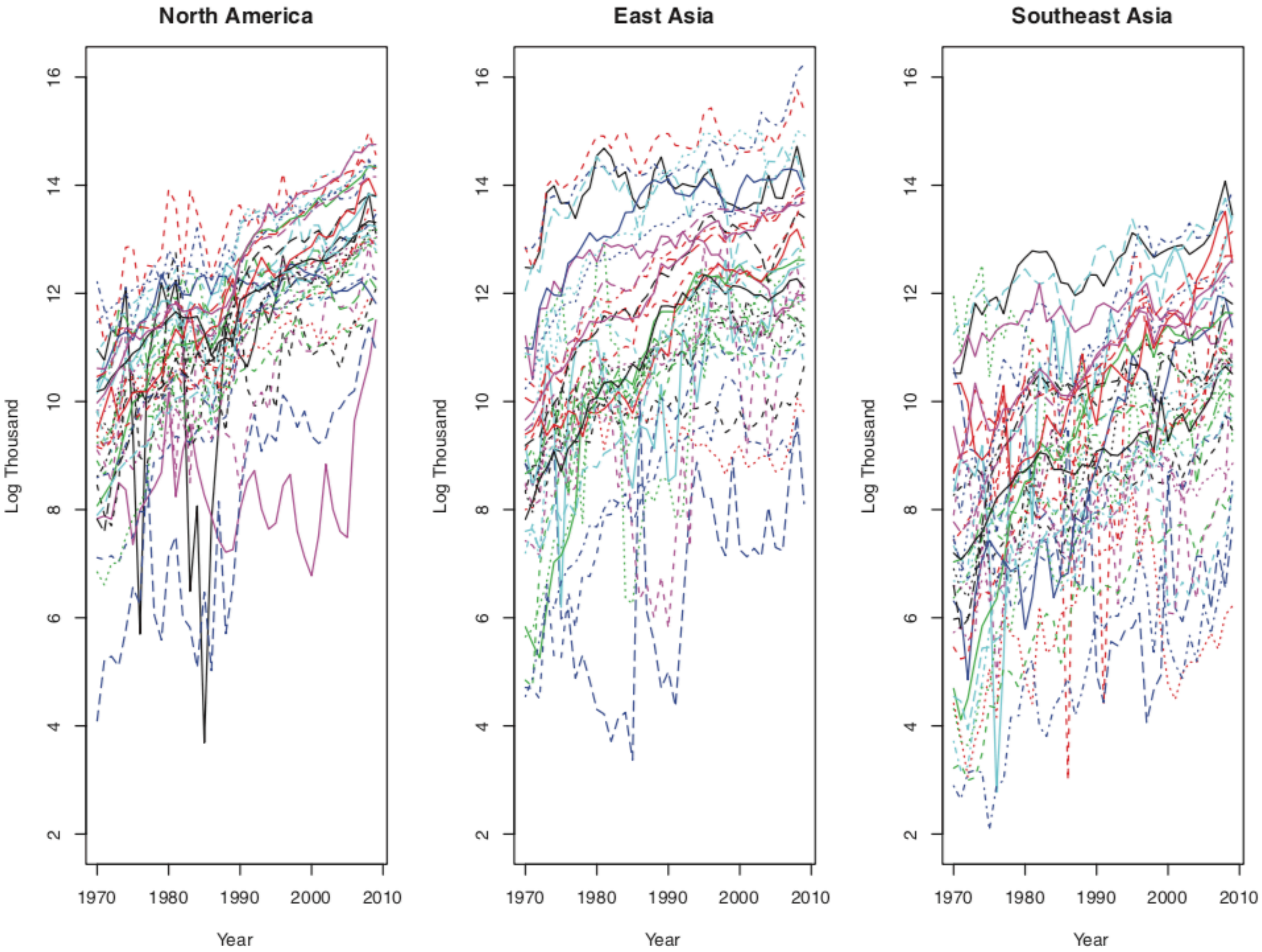
Application

- Data: Annual U.S. agricultural export data between 1970 and 2009 in thousands of U.S. dollars (40 observations).
- Each observation has the form:

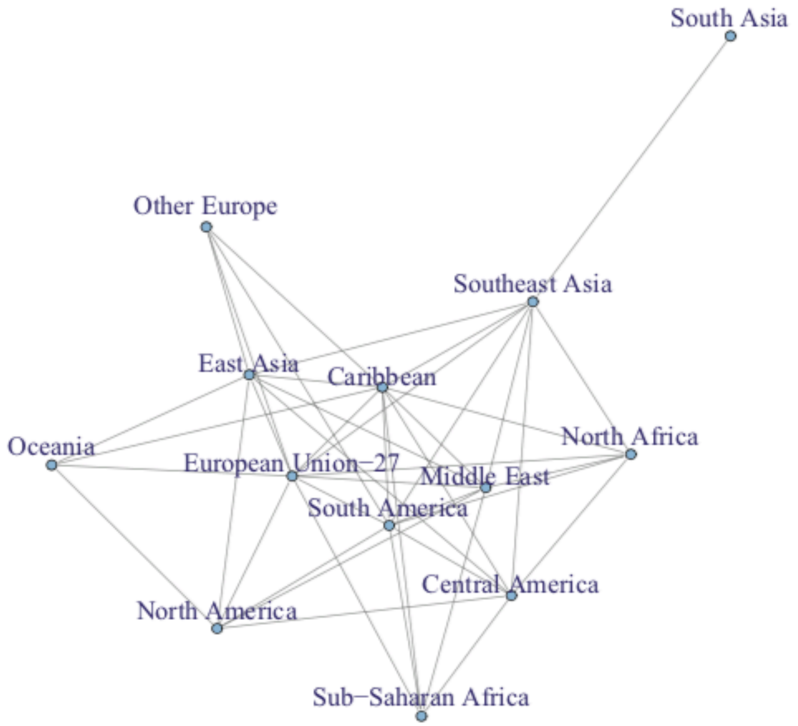
36 export items

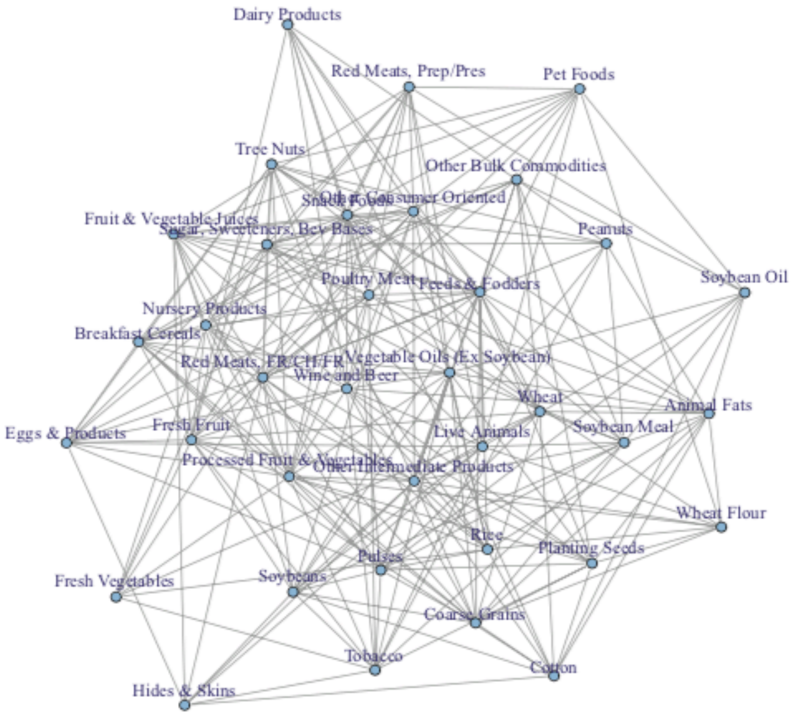
13 regions

	wheat flour	soybean oil	$\cdots$	fresh fruit	cotton
North America	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,35}$	$x_{1,36}$
Caribbean	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,35}$	$x_{2,36}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Africa	$x_{12,1}$	$x_{12,2}$	$\cdots$	$x_{12,35}$	$x_{12,36}$
Oceania	$x_{13,1}$	$x_{13,2}$	$\cdots$	$x_{13,35}$	$x_{13,36}$


North America
East Asia
Southeast Asia
Log Thousand
2
4
6
8
10
12
14
16
1970
1980
1990
2000
2010
Year

- We would like a graph for the regions and another for the export items:





- This can be achieved by using a graphical model on a **matrix-variate normal distribution** which encodes the structural information in row and column variables.

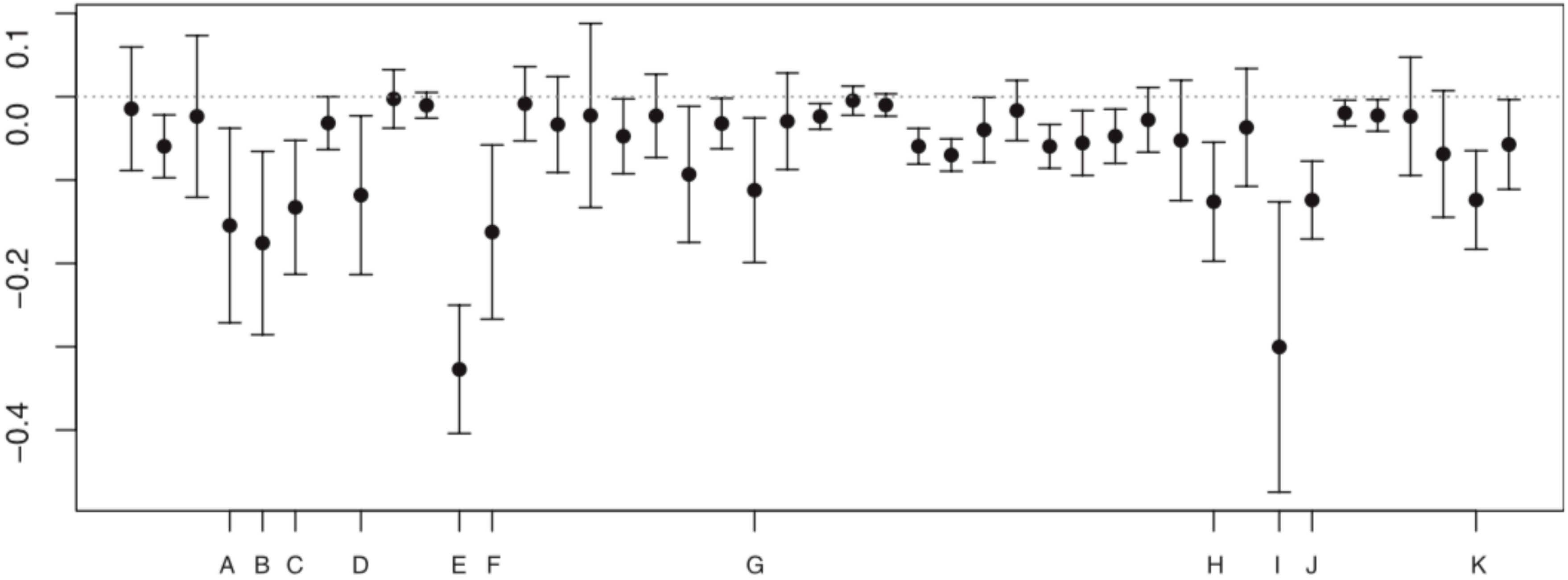


Figure 4. The estimates and the corresponding 95% confidence intervals for the 43 edges between the regions, where the largest 11 edges are marked. “A” is for CAR and SAM, “B” for CAM and SAM, “C” for NAM and EU, “D” for SAM and EU, “E” for EU and OE, “F” for NAM and EA, “G” for SAM and EU, “H” for SAM and SEA, “I” for EA and SEA, “J” for ME and SEA, and “K” for EU and OC, where EU denotes Europe Union, OE denotes Other Europe, EA denotes East Asia, SEA denotes Southeast Asia, CAM denotes Central America, NAM denotes North America, SAM denotes South America, CAR denotes Caribbean, and OC denotes Oceania.

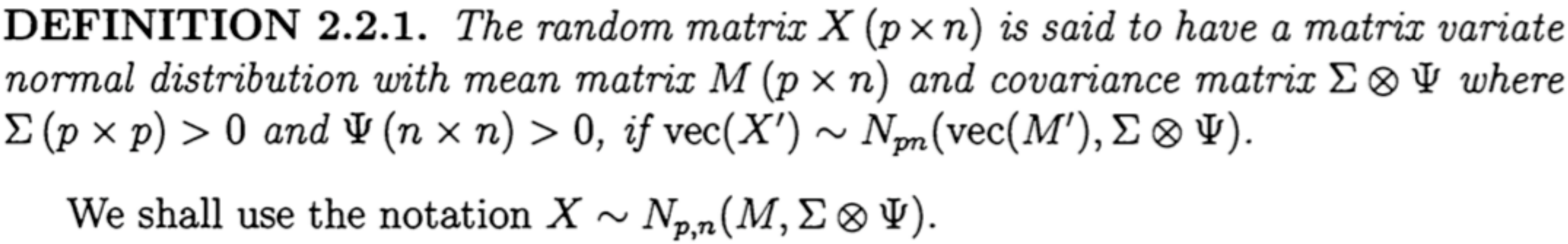
Basics: Matrix Variate Normal Distribution

Preliminaries: Normal Distributions

- $X \sim N(\mu, \sigma)$ has density: $(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},\ x \in \mathbb{R},$

- $\boldsymbol{X} = (X_1, \ldots, X_p) \sim N_p(\boldsymbol{\mu}, \Sigma)$ has density:

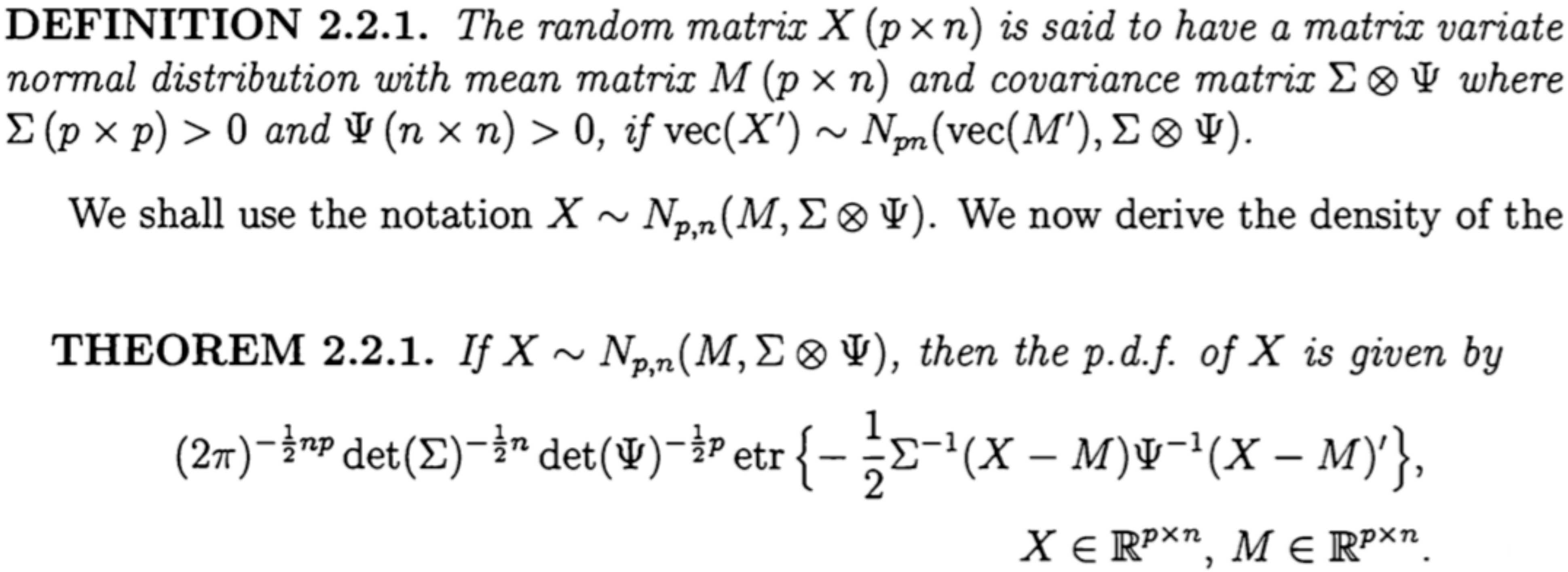
$$(2\pi)^{-\frac{1}{2}p} \det(\Sigma)^{-\frac{1}{2}} \operatorname{etr}\left\{-\frac{1}{2}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})'\right\},\ \boldsymbol{x} \in \mathbb{R}^p,\ \boldsymbol{\mu} \in \mathbb{R}^p,\ \Sigma > 0,$$



DEFINITION 2.2.1. *The random matrix* $X\,(p \times n)$ *is said to have a matrix variate normal distribution with mean matrix* $M\,(p \times n)$ *and covariance matrix* $\Sigma \otimes \Psi$ *where* $\Sigma\,(p \times p) > 0$ *and* $\Psi\,(n \times n) > 0$*, if* $\operatorname{vec}(X') \sim N_{pn}(\operatorname{vec}(M'), \Sigma \otimes \Psi)$.

We shall use the notation $X \sim N_{p,n}(M, \Sigma \otimes \Psi)$.

Matrix Variate Normal Distribution



DEFINITION 2.2.1. *The random matrix* $X\,(p \times n)$ *is said to have a matrix variate normal distribution with mean matrix* $M\,(p \times n)$ *and covariance matrix* $\Sigma \otimes \Psi$ *where* $\Sigma\,(p \times p) > 0$ *and* $\Psi\,(n \times n) > 0$, *if* $\operatorname{vec}(X') \sim N_{pn}(\operatorname{vec}(M'), \Sigma \otimes \Psi)$.

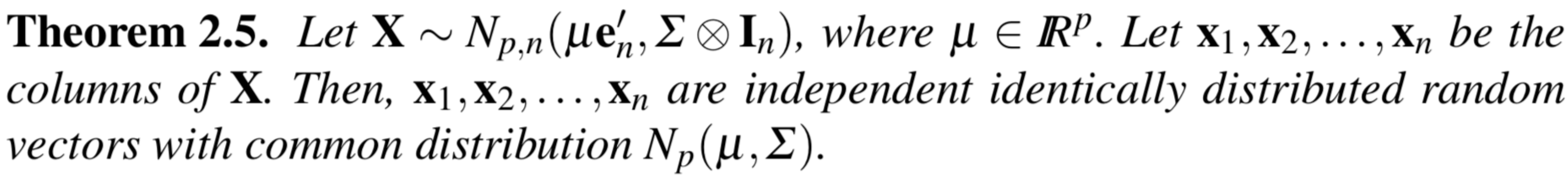
We shall use the notation $X \sim N_{p,n}(M, \Sigma \otimes \Psi)$. We now derive the density of the

THEOREM 2.2.1. *If* $X \sim N_{p,n}(M, \Sigma \otimes \Psi)$, *then the p.d.f. of* X *is given by*

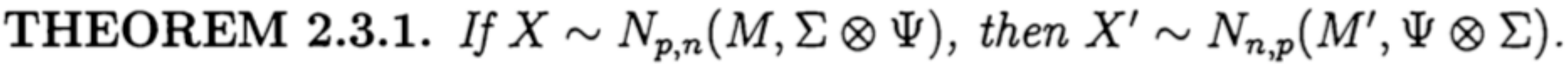
$$(2\pi)^{-\frac{1}{2}np} \det(\Sigma)^{-\frac{1}{2}n} \det(\Psi)^{-\frac{1}{2}p} \operatorname{etr}\left\{-\frac{1}{2}\Sigma^{-1}(X-M)\Psi^{-1}(X-M)'\right\},$$

$$X \in \mathbb{R}^{p\times n},\ M \in \mathbb{R}^{p\times n}.$$

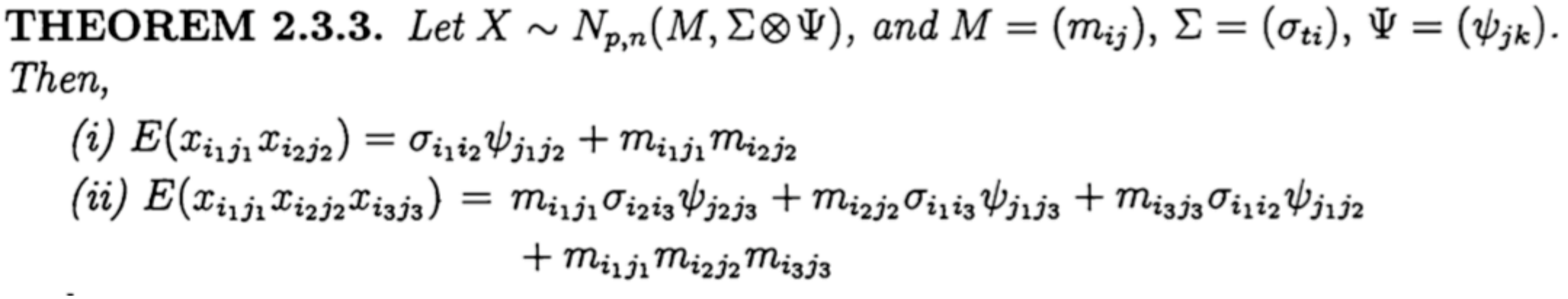
Some Properties of Matrix Variate Normal



Theorem 2.5. *Let* $\mathbf{X} \sim N_{p,n}(\mu \mathbf{e}_n', \Sigma \otimes \mathbf{I}_n)$*, where* $\mu \in \mathbb{R}^p$*. Let* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ *be the columns of* $\mathbf{X}$*. Then,* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ *are independent identically distributed random vectors with common distribution* $N_p(\mu, \Sigma)$*.*



THEOREM 2.3.1. *If* $X \sim N_{p,n}(M, \Sigma \otimes \Psi)$*, then* $X' \sim N_{n,p}(M', \Psi \otimes \Sigma)$*.*



THEOREM 2.3.3. *Let* $X \sim N_{p,n}(M, \Sigma \otimes \Psi)$*, and* $M = (m_{ij})$, $\Sigma = (\sigma_{ti})$, $\Psi = (\psi_{jk})$*. Then,*

(i) $E(x_{i_1 j_1} x_{i_2 j_2}) = \sigma_{i_1 i_2} \psi_{j_1 j_2} + m_{i_1 j_1} m_{i_2 j_2}$

(ii) $E(x_{i_1 j_1} x_{i_2 j_2} x_{i_3 j_3}) = m_{i_1 j_1} \sigma_{i_2 i_3} \psi_{j_2 j_3} + m_{i_2 j_2} \sigma_{i_1 i_3} \psi_{j_1 j_3} + m_{i_3 j_3} \sigma_{i_1 i_2} \psi_{j_1 j_2}$
$+ m_{i_1 j_1} m_{i_2 j_2} m_{i_3 j_3}$

Sparse Matrix Graphical Models

Sparse Matrix Graphical Models

- Let $\boldsymbol{X_1}, \dots, \boldsymbol{X_n} \sim \boldsymbol{N_{p,q}}(\boldsymbol{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$

where $\mathbf{M} \in \mathbb{R}^{p \times q}$ is the mean matrix, $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{q \times q}$

- WLOG assume $\mathbf{M} = \mathbf{0}$. Otherwise, subtract $\widehat{\mathrm{M}} = \sum X_i / n$

- Denote the precision matrices as: $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\omega_{ij})$ and $\boldsymbol{\Gamma} = \boldsymbol{\Psi}^{-1} = (\gamma_{ij})$
- The partial correlation between X_{ij} and X_{kl} is:

$$\rho_{ij,kl} = -\frac{\omega_{ik}}{\sqrt{\omega_{ii}\omega_{kk}}} \cdot \frac{\gamma_{jl}}{\sqrt{\gamma_{jj}\gamma_{ll}}}$$

- For identifiability, set $\omega_{11} = 1$

Sparse Matrix Graphical Models (SMGMs)

- The negative log-likelihood is:

$$\ell(\boldsymbol{\Omega}, \boldsymbol{\Gamma}) = -\frac{nq}{2} \log |\boldsymbol{\Omega}| - \frac{np}{2} \log |\boldsymbol{\Gamma}| + \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr} \left(\mathbf{X}_i \boldsymbol{\Gamma} \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Omega} \right)$$

- Sparse models for Ω and Σ can be obtained minimizing:

$$g(\boldsymbol{\Omega}, \boldsymbol{\Gamma}) = \frac{1}{npq} \sum_{i=1}^{n} \operatorname{tr} \left(\mathbf{X}_i \boldsymbol{\Gamma} \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Omega} \right) - \frac{1}{p} \log |\boldsymbol{\Omega}| - \frac{1}{q} \log |\boldsymbol{\Gamma}| + p_{\lambda_1}(\boldsymbol{\Omega}) + p_{\lambda_2}(\boldsymbol{\Gamma})$$

where $p_\lambda(\mathbf{A}) = \sum_{i \neq j} p_\lambda(|a_{ij}|)$ for a square matrix $\mathbf{A} = (a_{ij})$

- The article uses LASSO and SCAD penalties.

Penalties

For LASSO: $p_\lambda(s) = \lambda|s|$

For SCAD: $p'_\lambda(s) = \lambda\left\{I(s \le \lambda) + \frac{(3.7\lambda - s)_+}{2.7\lambda} I(s > \lambda)\right\}$

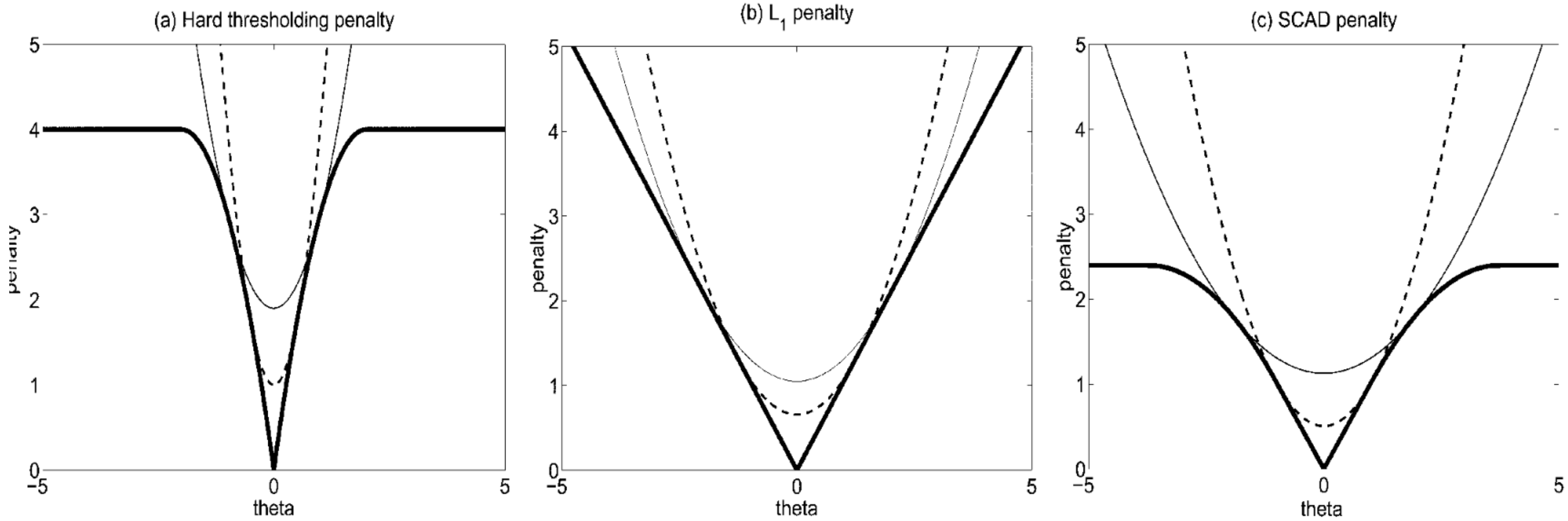


Figure 1. Three Penalty Functions $p_\lambda(\theta)$ and Their Quadratic Approximations. The values of λ are the same as those in Figure 5(c).

SMGM Implementation

$$g(\mathbf{\Omega}, \mathbf{\Gamma}) = \frac{1}{npq} \sum_{i=1}^{n} \mathrm{tr}\left(\mathbf{X}_i \mathbf{\Gamma} \mathbf{X}_i^{\mathrm{T}} \mathbf{\Omega}\right) - \frac{1}{p} \log |\mathbf{\Omega}| - \frac{1}{q} \log |\mathbf{\Gamma}| + p_{\lambda_1}(\mathbf{\Omega}) + p_{\lambda_2}(\mathbf{\Gamma})$$

- Let $\tilde{\mathbf{\Sigma}} = \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \mathbf{\Omega} \mathbf{X}_i / q$ and $\tilde{\mathbf{\Psi}} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{\Gamma} \mathbf{X}_i^{\mathrm{T}} / p$
- The penalized log-likelihood is not convex. But is conditionally convex if the penalty function is convex. When $\mathbf{\Omega}$ is fixed, solve:

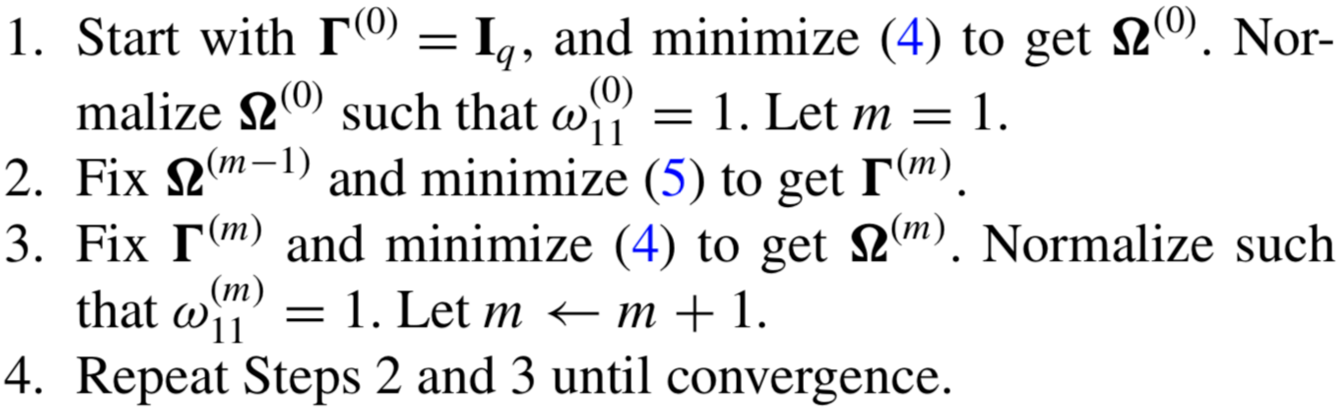
$$\min_{\Gamma} \frac{1}{q} \mathrm{tr}(\mathbf{\Gamma} \tilde{\mathbf{\Sigma}}) - \frac{1}{q} \log |\mathbf{\Gamma}| + p_{\lambda_2}(\mathbf{\Gamma})$$

- When $\mathbf{\Gamma}$ is fixed, solve:

$$\min_{\Omega} \frac{1}{p} \mathrm{tr}(\tilde{\mathbf{\Psi}} \mathbf{\Omega}) - \frac{1}{p} \log |\mathbf{\Omega}| + p_{\lambda_1}(\mathbf{\Omega})$$

SMGM Implementation

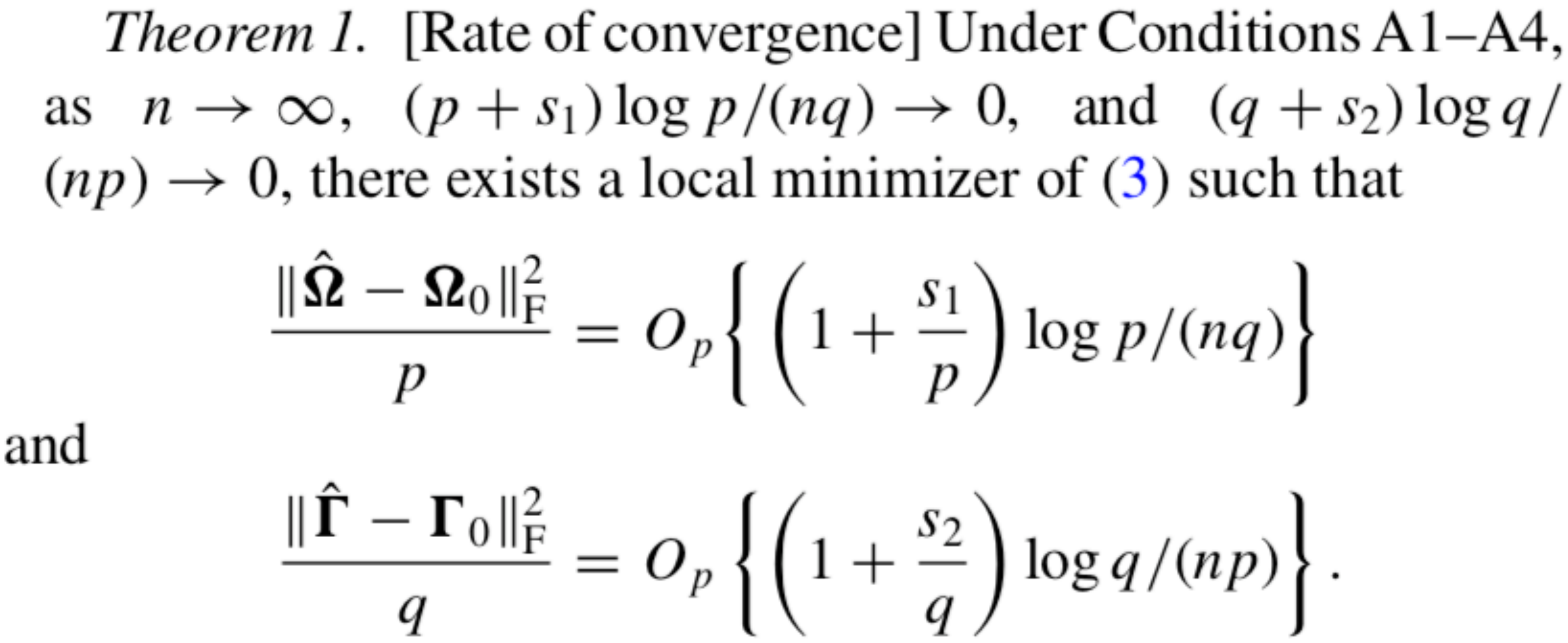
- The computational algorithm



1. Start with $\mathbf{\Gamma}^{(0)} = \mathbf{I}_q$, and minimize (4) to get $\mathbf{\Omega}^{(0)}$. Normalize $\mathbf{\Omega}^{(0)}$ such that $\omega_{11}^{(0)} = 1$. Let $m = 1$.
2. Fix $\mathbf{\Omega}^{(m-1)}$ and minimize (5) to get $\mathbf{\Gamma}^{(m)}$.
3. Fix $\mathbf{\Gamma}^{(m)}$ and minimize (4) to get $\mathbf{\Omega}^{(m)}$. Normalize such that $\omega_{11}^{(m)} = 1$. Let $m \leftarrow m + 1$.
4. Repeat Steps 2 and 3 until convergence.

- For $n < \min(p, q)$: The algorithm converges to a local stationary point of $g(\mathbf{\Omega}, \mathbf{\Gamma})$. But there is no guarantee on the global minimum.
- For $n > \mathrm{m}ax(p, q)$: MLE exists and is unique. Moreover the algorithm guarantees global optimal solution for LASSO or ridge.

Asymptotics



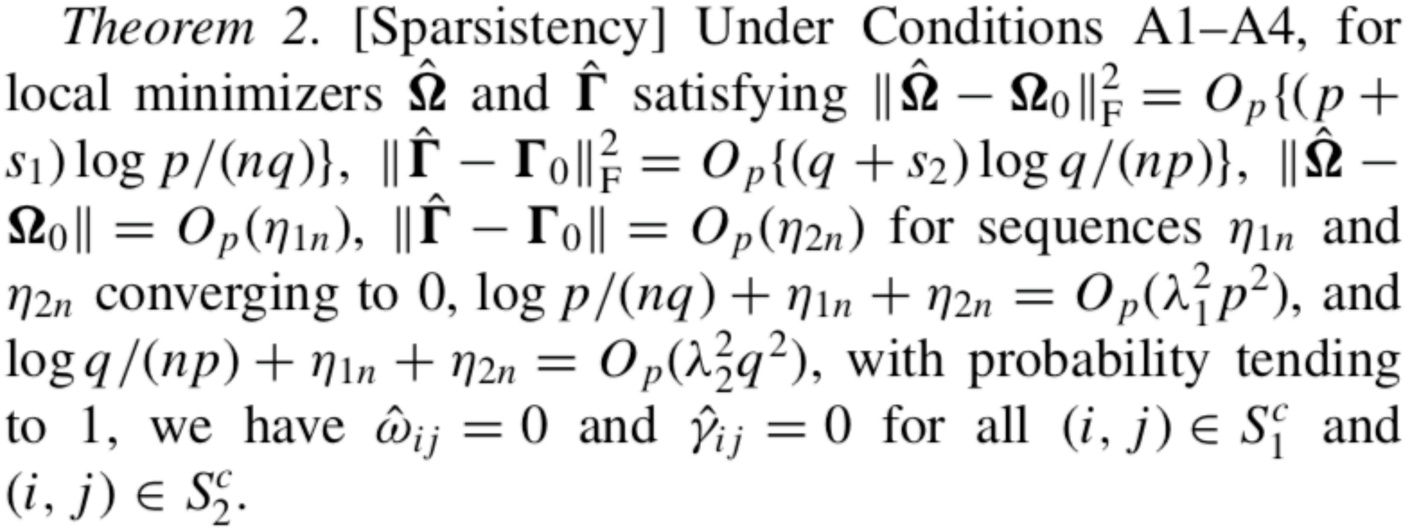
Theorem 1. [Rate of convergence] Under Conditions A1–A4, as $n \to \infty$, $(p + s_1)\log p/(nq) \to 0$, and $(q + s_2)\log q/(np) \to 0$, there exists a local minimizer of (3) such that

$$\frac{\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_{\mathrm{F}}^2}{p} = O_p\left\{\left(1 + \frac{s_1}{p}\right)\log p/(nq)\right\}$$

and

$$\frac{\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\|_{\mathrm{F}}^2}{q} = O_p\left\{\left(1 + \frac{s_2}{q}\right)\log q/(np)\right\}.$$

Asymptotics

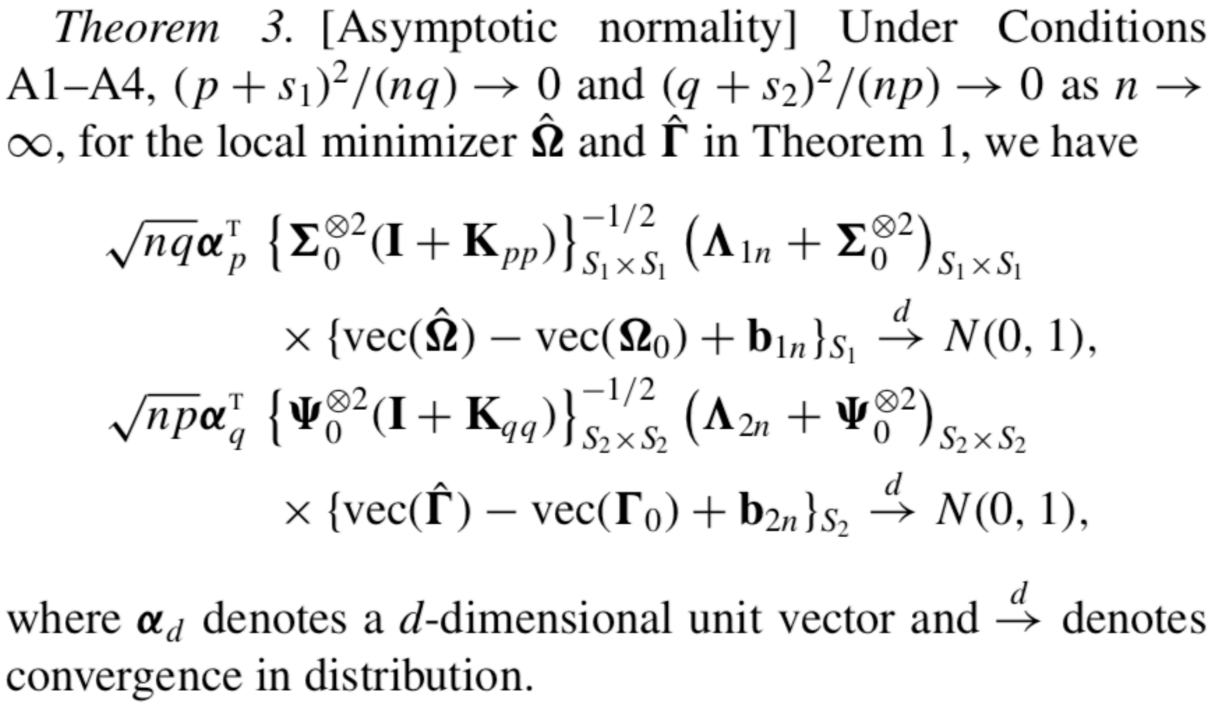


Theorem 2. [Sparsistency] Under Conditions A1–A4, for local minimizers $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{\Gamma}}$ satisfying $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_{\mathrm{F}}^2 = O_p\{(p + s_1)\log p/(nq)\}$, $\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\|_{\mathrm{F}}^2 = O_p\{(q + s_2)\log q/(np)\}$, $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\| = O_p(\eta_{1n})$, $\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\| = O_p(\eta_{2n})$ for sequences η_{1n} and η_{2n} converging to 0, $\log p/(nq) + \eta_{1n} + \eta_{2n} = O_p(\lambda_1^2 p^2)$, and $\log q/(np) + \eta_{1n} + \eta_{2n} = O_p(\lambda_2^2 q^2)$, with probability tending to 1, we have $\hat{\omega}_{ij} = 0$ and $\hat{\gamma}_{ij} = 0$ for all $(i, j) \in S_1^c$ and $(i, j) \in S_2^c$.

- Sparsistency when the off-diagonal nonzeros of $\mathbf{\Omega}_0$ and $\mathbf{\Gamma}_0$ are:
  - For LASSO: at most $O(p)$ and $O(q)$
  - For SCAD: at most $O(p^2)$ and $O(q^2)$

Asymptotics

Fast rates of convergence $\sqrt{nq}$ and $\sqrt{np}$ for estimating the precision matrices



Theorem 3. [Asymptotic normality] Under Conditions A1–A4, $(p+s_1)^2/(nq) \to 0$ and $(q+s_2)^2/(np) \to 0$ as $n \to \infty$, for the local minimizer $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{\Gamma}}$ in Theorem 1, we have

$$\begin{aligned}
&\sqrt{nq}\boldsymbol{\alpha}_p^{\mathrm{T}} \left\{\mathbf{\Sigma}_0^{\otimes 2}(\mathbf{I}+\mathbf{K}_{pp})\right\}_{S_1\times S_1}^{-1/2} \left(\mathbf{\Lambda}_{1n}+\mathbf{\Sigma}_0^{\otimes 2}\right)_{S_1\times S_1} \\
&\qquad \times \{\mathrm{vec}(\hat{\mathbf{\Omega}}) - \mathrm{vec}(\mathbf{\Omega}_0) + \mathbf{b}_{1n}\}_{S_1} \xrightarrow{d} N(0,1), \\
&\sqrt{np}\boldsymbol{\alpha}_q^{\mathrm{T}} \left\{\mathbf{\Psi}_0^{\otimes 2}(\mathbf{I}+\mathbf{K}_{qq})\right\}_{S_2\times S_2}^{-1/2} \left(\mathbf{\Lambda}_{2n}+\mathbf{\Psi}_0^{\otimes 2}\right)_{S_2\times S_2} \\
&\qquad \times \{\mathrm{vec}(\hat{\mathbf{\Gamma}}) - \mathrm{vec}(\mathbf{\Gamma}_0) + \mathbf{b}_{2n}\}_{S_2} \xrightarrow{d} N(0,1),
\end{aligned}$$

where $\boldsymbol{\alpha}_d$ denotes a d-dimensional unit vector and $\xrightarrow{d}$ denotes convergence in distribution.