Minimax Estimation of Large Precision Matrices with Bandable Cholesky Factor

Zhao Ren

University of Pittsburgh, USA

February 6th, 2018 @NUS

Outline

Introduction: Covariance/precision matrices estimation

- Precision matrices are the inverse of covariance matrices. They are important in many statistical methods, such as PCA, LDA/QDA, regression, clustering analysis and graphical models.
- In high-dimensional setting, the sample covariance matrix is not consistent. (e.g., [Johnstone, 2001]).
- Structural assumptions on matrices are needed in order to overcome the difficulty due to high-dimensionality.

Introduction: Structures

- "Sparsity"
  - Unordered: sparse covariance/precision matrices;
  - Ordered: bandable covariance, precision with bandable Cholesky factor.
- More complicated: Spiked covariance matrices, Covariance with tensor product, latent graphical models, etc.

Introduction: Sparsity Structures

On the covariance matrix:

- sparse: [d'Aspremont et al., 2008], [Cai and Zhou, 2012]...
- bandable: [Bickel and Levina, 2008a], [Bickel and Levina, 2008b], [Cai et al., 2010]...

On the precision matrix:

- sparse: [Yuan and Lin, 2007], [Meinshausen and Bühlmann, 2006], [Ren et al., 2015],...
- "bandable": [Bickel and Levina, 2008b], [Lee and Lee, 2017]...

Introduction: Sparsity Structures

Minimax framework:

On the covariance matrix:

- sparse: [d'Aspremont et al., 2008], [Cai and Zhou, 2012]...
- bandable: [Bickel and Levina, 2008a], [Bickel and Levina, 2008b], [Cai et al., 2010]...

On the precision matrix:

- sparse: [Yuan and Lin, 2007], [Meinshausen and Bühlmann, 2006], [Ren et al., 2015]...
- "bandable": [Bickel and Levina, 2008b], [Lee and Lee, 2017]...

Introduction: Sparsity Structures

Minimax framework:

On the covariance matrix:

- sparse: [d'Aspremont et al., 2008], [Cai and Zhou, 2012]...
- bandable: [Bickel and Levina, 2008a], [Bickel and Levina, 2008b], [Cai et al., 2010]...

On the precision matrix:

- sparse: [Yuan and Lin, 2007], [Meinshausen and Bühlmann, 2006], [Ren et al., 2015]...
- "bandable": [Bickel and Levina, 2008b], [Lee and Lee, 2017]...

Minimax Estimation of Large Precision Matrices with Bandable Cholesky Factor

Cholesky decomposition of precision matrices

Build the connection between the regression and precision matrices:
Assume $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$ is the p-variate random vector,

Auto-regression:

$$
\begin{aligned}
X_1 &= 0 + \epsilon_1 \\
X_2 &= a_{21} X_1 + \epsilon_2 \\
X_3 &= a_{32} X_2 + a_{31} X_1 + \epsilon_3 \\
&\dots \\
X_p &= a_{p(p-1)} X_{p-1} + a_{p(p-2)} X_{p-2} + \cdots + a_{p1} X_1 + \epsilon_p
\end{aligned}
$$

Cholesky decomposition of precision matrices

Build the connection between the regression and precision matrices:
Assume $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ is the p-variate random vector,

Rewrite it as:

$$
\begin{aligned}
X_1 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad &= \epsilon_1 \\
-a_{21}X_1 + \; X_2 \qquad\qquad\qquad\qquad\qquad\qquad &= \epsilon_2 \\
-a_{31}X_1 - a_{32}X_2 + \; X_3 \qquad\qquad\qquad\qquad &= \epsilon_3 \\
\ldots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad &\ldots \\
-a_{p1}X_1 - a_{p2}X_2 - a_{p3}X_3 \cdots - a_{p(p-1)}X_{p-1} + X_p &= \epsilon_p
\end{aligned}
$$

Cholesky decomposition of precision matrices

The matrix form:

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -a_{21} & 1 & 0 & \dots & 0 \\ -a_{31} & -a_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ -a_{p1} & -a_{p2} & -a_{p3} & \dots & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

$$(I - A)\mathbf{X} = \boldsymbol{\epsilon}$$

$$\Sigma = (I - A)^{-1} D (I - A)^{-T}$$

$$\Omega = (I - A)^{T} D^{-1} (I - A)$$

where A is a lower triangular matrix with zero diagonals, D is a diagonal matrix.

Cholesky decomposition of precision matrices - Example

Example

The autoregressive model in time series: AR(1)

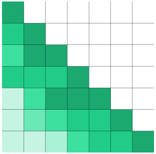
$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -a_{21} & 1 & 0 & \dots & 0 \\ 0 & -a_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

In AR(k) model, A is a k-banded matrix.

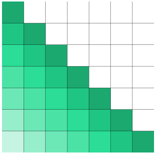
Bandable structures on the Cholesky factors

[Bickel and Levina, 2008b, Cai et al., 2010] proposed two different bandable structures:

$$\max_i \sum_{j<i-k} |a_{ij}| < Mk^{-\alpha}, \quad \forall 1 \leq k \leq p$$



$$|a_{ij}| < M(i-j)^{-\alpha-1}, \quad \forall 1 \leq j \leq i-1$$



Parameter spaces

We consider two bandable structures on the Cholesky factors of precision matrices mentioned above:

Assume that $\Omega = (I - A)^T D^{-1}(I - A)$ For $M > 0$, $\eta > 1$,

$$\mathcal{P}_\alpha(\eta, M) = \Big\{\Omega : \quad \eta^{-1} \le \lambda_{min}(\Omega) \le \lambda_{max}(\Omega) < \eta, \\ \max_i \sum_{j<i-k} |a_{ij}| < Mk^{-\alpha}, \quad \forall 1 \le k \le p\Big\},$$

$$\mathcal{Q}_\alpha(\eta, M) = \Big\{\Omega : \quad \eta^{-1} \le \lambda_{min}(\Omega) \le \lambda_{max}(\Omega) < \eta, \\ |a_{ij}| < M(i-j)^{-\alpha-1}, \quad \forall 1 \le j \le i-1\Big\}.$$

Remark: $\mathcal{Q}_\alpha(\eta, \alpha M) \subset \mathcal{P}_\alpha(\eta, M)$.

A minimax decision framework

- Minimax framework is one way to evaluate the performance of estimators within a given parameter space.
- Given a parameter space Θ and a loss function $L(\cdot,\cdot)$, one is looking for the optimal rate of convergence

$$R^* \asymp \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E} L(\theta, \hat{\theta}).$$

- We consider Operator norm and Frobenius norm in this talk
  - Operator norm:

$$\|X\|_{\mathrm{op}} = \sup_{a \neq 0} \{ \frac{\|Xa\|_2}{\|a\|_2} \}$$

    It is the largest singular value of the matrix.
  - Frobenius norm:

$$\|X\|_{\mathrm{F}} = (\sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij}^2)^{\frac{1}{2}}.$$

    It treats the matrix as a long vector, it is the L_2 norm of that vector.

Our Goals:

Given n i.i.d samples, we consider the minimax risks in estimating the precision matrix Ω of $\mathbf{X}$, over two parameter spaces $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$, under Operator norm and Frobenius norm.

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \qquad \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2$$

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\frac{1}{p}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \qquad \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\frac{1}{p}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2$$

Why the rate optimality was not developed?

A striking phenomenon

- Intuitively, one would expect the same minimax rates of convergence under the operator norm between estimating bandable covariance matrices and precision matrices with bandable Cholesky factor.

- [Cai et al., 2010] established the optimal rate of convergence $\mathbb{E}\|\tilde{\Sigma} - \Sigma\|_{\text{op}}^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$ for bandable covariance matrices $\Sigma = \Omega^{-1} = [\sigma_{ij}]_{p \times p}$ such that $\max_i \sum_{|j-i|>k} |\sigma_{ij}| < Mk^{-\alpha}$, $k \in [p]$.

- We show a surprising result: estimation over $\mathcal{P}_\alpha(\eta, M)$ is a much harder task than that over bandable covariance matrices.

Existing procedures

- Almost all existing approaches reply on an intermediate estimator $\hat{A}$ of A via regressions (i.e., estimator of each $\mathbf{a}_i$). For example, [Wu and Pourahmadi, 2003], [Huang et. al, 2006], [Levina, Rothman and Zhu, 2008], [Bickle and Levina (2008b)], [Fan, Xue and Zou, 2016], etc.

- Analysis relies on bounding $\max_i \|\mathbf{a}_i - \hat{\mathbf{a}}_i\|$ in order to bound $\|\hat{A} - A\|_{\mathrm{op}}^2$.

- The analysis above usually is not sharp. (e.g., [Cai et al., 2010] for bandable covariance matrix estimation)

Existing procedures - Bickle and Levina (2008b)

Since the Cholesky factors of Ω has the bandable structure, Bickle and Levina approximated A by the k-banded matrix A_k.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ -a_{21} & 1 & 0 & 0 & 0 & \dots & 0 \\ -a_{31} & -a_{32} & 1 & 0 & 0 & \dots & 0 \\ -a_{41} & -a_{42} & -a_{43} & 1 & 0 & \dots & 0 \\ -a_{51} & -a_{52} & -a_{53} & -a_{54} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ -a_{p1} & -a_{p2} & \dots & -a_{p(p-3)} & -a_{p(p-2)} & -a_{p(p-1)} & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ \vdots \\ \vdots \\ X_p \end{bmatrix}$$

$$X_i = \mathbf{a}_i X_{1:i-1} + \epsilon_i \quad var(\epsilon_i) = d_i$$

Existing procedures - Bickle and Levina (2008b)

Since the Cholesky factors of Ω has the bandable structure, Bickle and Levina approximate A by the k-banded matrix B_k.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ -b_{21} & 1 & 0 & 0 & 0 & \dots & 0 \\ -b_{31} & -b_{32} & 1 & 0 & 0 & \dots & 0 \\ 0 & -b_{42} & -b_{43} & 1 & 0 & \dots & 0 \\ 0 & 0 & -b_{53} & -b_{54} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 0 & -b_{p(p-2)} & -b_{p(p-1)} & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ \vdots \\ \vdots \\ X_p \end{bmatrix}$$

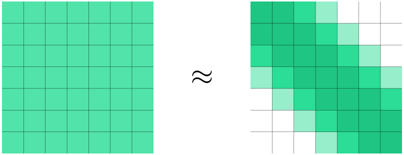
$$X_i = \mathbf{b}_i X_{i-k:i-1} + \delta_i \quad var(\delta_i) = f_i$$

Minimax risk under operator norm

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \qquad \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2$$

Estimation procedure: Motivation I

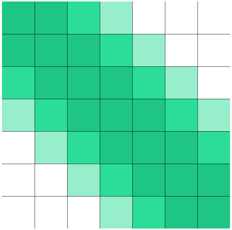
The bandable structure on the Cholesky factors implies "certain" bandable structure on the precision matrix.



What we have learned from estimating bandable covariance matrices [Cai et al., 2010]?

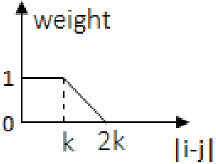
Estimation procedure: Motivation II

For bandable covariance $\Sigma = (\sigma_{ij})$, a direct target is a tapered population covariance with bandwidth k [Cai et al., 2010]:

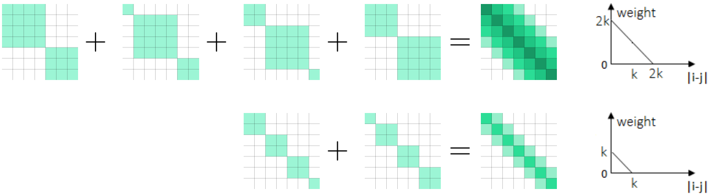
$\Sigma_k^{tap} =$ 

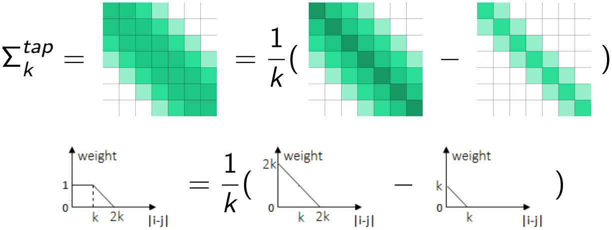
The tapered population covariance:

$$\sigma_{ij}^{tap} = \sigma_{ij} w_{ij}$$



Estimation procedure: Motivation II





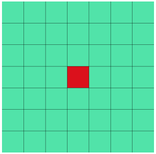
Estimation procedure: Motivation

- The core analysis relies on a rate-optimal estimator of each principal submatrix of Σ of smaller size k under operator norm: local sample covariance of size k.

- How should we estimate each principal submatrix of Ω of smaller size k? Inversing local sample covariance of size k is NOT optimal?

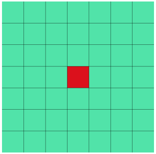
Estimation procedure: Local cropping estimator

Target: each principal submatrix of the precision matrix, $\Omega^{loc}_{m,k}$



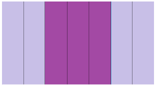
Estimation procedure: Local cropping estimator

Target: each principal submatrix of the precision matrix, $\Omega^{loc}_{m,k}$

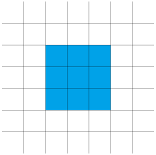


Estimator: $\hat{\Omega}^{loc}_{m,k}$

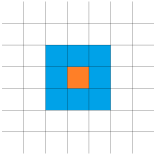
1.Collect the observation of $X_{m-k:m+2k-1}$:



2.Calculate the sample precision matrix:



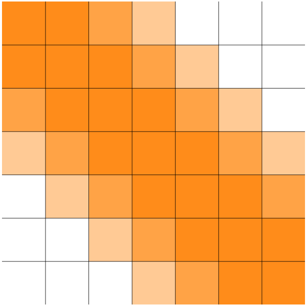
3.Crop its central part as the local estimator:



Estimation procedure: Local cropping estimator

Our final estimator is

$$\hat{\Omega}_k = \frac{1}{k}\left(\sum_{m=2-2k}^{p} \hat{\Omega}^{loc}_{m,2k} - \sum_{m=2-k}^{p} \hat{\Omega}^{loc}_{m,k}\right)$$



Upper bound: Analysis

- The local cropping estimator can be written as a sum of many principal submatrix estimators.
- There is natural bias and variance trade off, when picking optimal bandwidth k.

$$\text{risk} = \text{variance} + \text{bias I} + \text{bias II}$$

- variance is due to $\hat{\Omega}^{loc}_{m,k} - \mathbb{E}\hat{\Omega}^{loc}_{m,k}$; bias I is due to $\mathbb{E}\hat{\Omega}^{loc}_{m,k} - \Omega^{loc}_{m,k}$.
- bias II is due to $\Omega - \Omega^{tap}_{k}$.

Remark: In constrast, the analysis of bandable covariance only has one bias term.

Upper bound - Variance

- The variance is controlled by the maximum variance among all principal submatrices estimators.

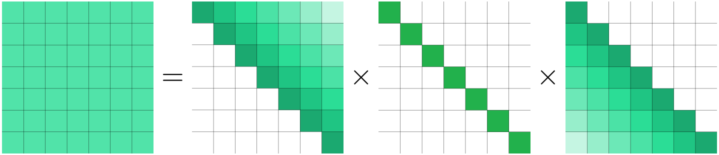
- By Bonferroni correction:

$$\max_m \mathbb{E}\|\hat{\Omega}^{loc}_{m,k} - \mathbb{E}\hat{\Omega}^{loc}_{m,k}\|^2_{\mathrm{op}} \leq C\frac{\log p + k}{n}.$$

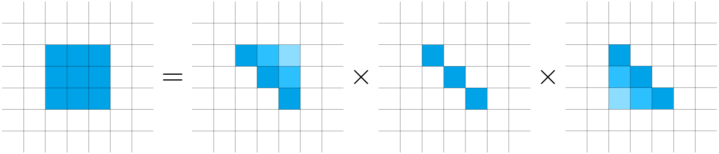
Upper bound - bias I over $\mathcal{P}_\alpha(\eta, M)$

The bias:

the Cholesky decomposition of the precision matrix:



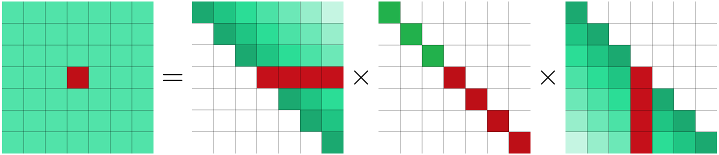
the Cholesky decomposition of the $3k$-precision matrix:



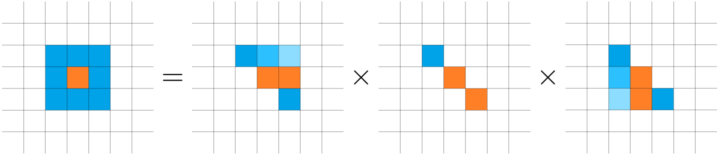
Upper bound - bias I over $\mathcal{P}_\alpha(\eta, M)$

The bias: $k^{1-2\alpha}$

the Cholesky decomposition of the precision matrix:



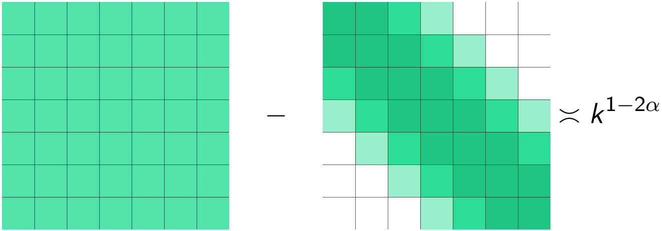
the Cholesky decomposition of the $3k$-precision matrix:



The bias: $k^{1-2\alpha}$ $k^{-2\alpha}$ $k^{1-2\alpha}$

Upper bound - bias II over $\mathcal{P}_\alpha(\eta, M)$

The bias of the entire matrix:



Remark: The proof is based on the block-wise analysis.

Upper bound over $\mathcal{P}_\alpha(\eta, M)$

The upper bound of the estimator:

- The variance: $\frac{\log p + k}{n}$.
- The bias I: $k^{1-2\alpha}$.
- The bias II: $k^{1-2\alpha}$.

Combining the above together, we find the upper bound of the estimator:

$$\sup_{\mathcal{P}_\alpha(\eta, M)} \mathbb{E}\|\hat{\Omega}_k - \Omega\|_{\mathrm{op}}^2 \leq Ck^{1-2\alpha} + C\frac{\log p + k}{n}.$$

Choose $k = n^{\frac{1}{2\alpha}}$, we have

$$\sup_{\mathcal{P}_\alpha(\eta, M)} \mathbb{E}\|\hat{\Omega}_k - \Omega\|_{\mathrm{op}}^2 \leq Cn^{-\frac{1-2\alpha}{2\alpha}} + C\frac{\log p}{n}.$$

Upper bound over $\mathcal{Q}_\alpha(\eta, M)$

The upper bound of the estimator:

- The variance: $\frac{\log p + k}{n}$.
- The bias I: $k^{-2\alpha}$.
- The bias II: $k^{-2\alpha}$.

Combining the above together, we find the upper bound of the estimator:

$$\sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\hat{\Omega}_k - \Omega\|_{\mathrm{op}}^2 \le Ck^{-2\alpha} + C\frac{\log p + k}{n}.$$

Choose $k = n^{\frac{1}{2\alpha+1}}$, we have

$$\sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\hat{\Omega}_k - \Omega\|_{\mathrm{op}}^2 \le Cn^{-\frac{2\alpha}{2\alpha+1}} + C\frac{\log p}{n}.$$

Lower bound

- Lower bound of the convergence rate characterize the difficulty of the estimation problem.
- The basic strategy is to select finite points in the parameter space, and then "reduce" it to a testing question.
- The difference between $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ is established by constructing the corresponding (different) minimax lower bounds: Assouad's Lemma.

Lower bound - construction in $\mathcal{Q}_\alpha(\eta, M)$

$$\mathcal{P}_1 = \left\{\Omega(\theta) : \Omega(\theta) = (I_p - A(\theta))^T (I_p - A(\theta)), \theta \in \Theta\right\}$$

where $$A(\theta) = \begin{bmatrix} 0_{k\times k} & 0_{k\times k} & 0_{k\times(p-2k)} \\ \boxed{\begin{matrix} (nk)^{-\frac{1}{2}} & \dots & (nk)^{-\frac{1}{2}} & (nk)^{-\frac{1}{2}} \\ (nk)^{-\frac{1}{2}} & \dots & (nk)^{-\frac{1}{2}} & (nk)^{-\frac{1}{2}} \\ \dots & \dots & \dots & \dots \\ (nk)^{-\frac{1}{2}} & \dots & (nk)^{-\frac{1}{2}} & (nk)^{-\frac{1}{2}} \end{matrix}} & 0_{k\times k} & 0_{k\times(p-2k)} \\ 0_{(p-2k)\times k} & 0_{(p-2k)\times k} & 0_{(p-2k)^2} \end{bmatrix}$$

where $\Theta = \{0,1\}^k$. $k = n^{\frac{1}{2\alpha+1}}$.
The lower bound over the subset $\mathcal{P}_1$ is:

$$\sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq \sup_{\mathcal{P}_1} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq Cn^{-\frac{2\alpha}{2\alpha+1}}$$

Lower bound - construction in $\mathcal{P}_\alpha(\eta, M)$

$$\mathcal{P}_2 = \left\{ \Omega(\theta) : \Omega(\theta) = (I_p - A(\theta))^T (I_p - A(\theta)), \theta \in \Theta \right\}$$

$$\text{where} \quad A(\theta) = \begin{bmatrix} 0_{k\times k} & 0_{k\times k} & 0_{k\times(p-2k)} \\ \boxed{\begin{matrix} 0 & \dots & 0 & n^{-\frac{1}{2}} \\ 0 & \dots & 0 & n^{-\frac{1}{2}} \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & n^{-\frac{1}{2}} \end{matrix}} & 0_{k\times k} & 0_{k\times(p-2k)} \\ 0_{(p-2k)\times k} & 0_{(p-2k)\times k} & 0_{(p-2k)^2} \end{bmatrix}$$

where $\Theta = \{0, 1\}^k$, $k = n^{\frac{1}{2\alpha}}$.

The lower bound over the subset $\mathcal{P}_2$ is:

$$\sup_{\mathcal{P}_\alpha(\eta, M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\rm op}^2 \geq \sup_{\mathcal{P}_2} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\rm op}^2 \geq Cn^{-\frac{2\alpha-1}{2\alpha}}$$

Main results: Minimax risk over $\mathcal{P}_\alpha(\eta, M)$

Theorem 1 (Minimax risk over $\mathcal{P}_\alpha(\eta, M)$)

The minimax risk of the precision matrix Ω with $\alpha > \frac{1}{2}$ over $\mathcal{P}_\alpha(\eta, M)$ satisfies

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \asymp n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}$$

this rate can be achieved by the local cropping estimator.

Remark: When $\alpha \leq 1/2$, there is NO consistent estimator for most settings!

Main results: Minimax risk over $\mathcal{Q}_\alpha(\eta, M)$

Theorem 2 (Minimax risk over $\mathcal{Q}_\alpha(\eta, M)$)

The minimax risk of the precision matrix Ω over $\mathcal{Q}_\alpha(\eta, M)$ satisfies

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$$

this rate can be achieved by the local cropping estimator.

Remark: The local cropping estimator is consistent as long as $\alpha > 0$.

Remark: The convergence rate of the banding estimator proposed by [Bickel and Levina, 2008b] is $(n/\log p)^{-\frac{2\alpha}{2\alpha+2}}$, which is sub-optimal.

Simulation studies in $\mathcal{Q}_\alpha(\eta, M)$

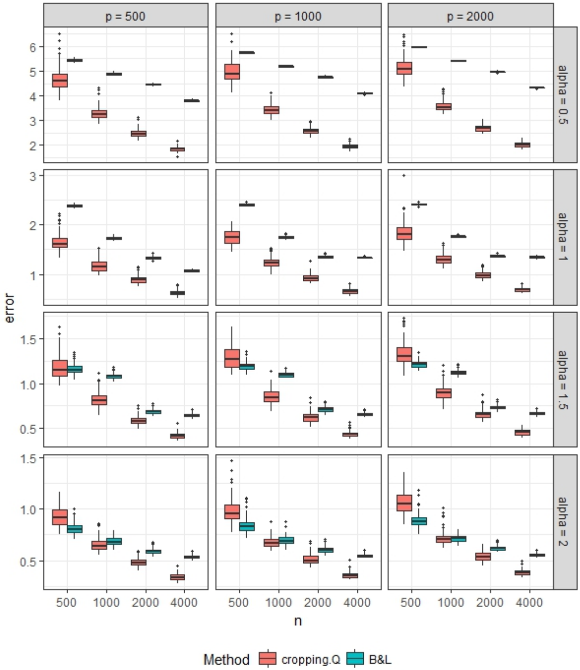
- Consider the precision matrix in the following form:

$$\Omega = (I - A)^T D^{-1} (I - A), \quad A = [a_{ij}]_{p \times p}, \quad D = I_p$$

where $a_{ij} = -(i-j)^{-\alpha-1}$ when $i > j$; otherwise $a_{ij} = 0$.

- cropping Q: The local cropping estimator with bandwidth $k = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$.

- B&L: The banding estimator proposed in [Bickel and Levina, 2008a] with bandwidth $k = \lfloor (n/\log p)^{1/(2\alpha+2)} \rfloor$.


p = 500
p = 1000
p = 2000
alpha = 0.5
alpha = 1
alpha = 1.5
alpha = 2
error
n
500 1000 2000 4000
Method cropping.Q B&L

Simulation studies in $\mathcal{P}_\alpha(\eta, M)$

- Consider the precision matrix in $\mathcal{P}_\alpha(\eta, M)$ but not always in $\mathcal{Q}_\alpha(\eta, M)$:

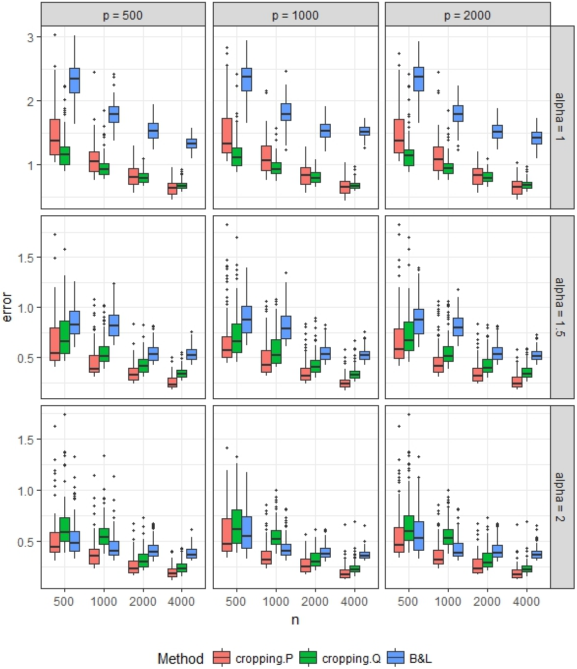
$$\Omega = (I - A)^T D^{-1} (I - A), \quad A = [a_{ij}]_{p \times p}, \quad D = I_p$$

where $a_{i1} = -2(i-1)^{-\alpha}$ when $2 \le i \le p$; otherwise $a_{ij} = 0$.

- cropping P: The local cropping estimator with optimal bandwidth $k = \lfloor n^{\frac{1}{2\alpha}} \rfloor$.

- cropping Q: The local cropping estimator with sub-optimal bandwidth $k = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$.

- B&L: The banding estimator proposed in [Bickel and Levina, 2008a] with bandwidth $k = \lfloor (n/\log p)^{1/(2\alpha+2)} \rfloor$.


p = 500
p = 1000
p = 2000
alpha = 1
alpha = 1.5
alpha = 2
error
n
500 1000 2000 4000
Method cropping.P cropping.Q B&L

Adaptive procedure

- Lepski's method: a popular data-driven procedure in many nonparametric estimation problems.
- Our adaptive (to the knowledge of α) procedure: With a discrete set of bandwidths $\mathcal{H} = \{1, \ldots, n/\log p\}$, we select $\hat{k}$ by

$$\hat{k} = \min\left\{k \in \mathcal{H} : \|\hat{\Omega}_k - \hat{\Omega}_l\|_{\mathrm{op}}^2 \le C\frac{l + \log p}{n} \text{ for all } l \ge k\right\}$$

- Main results:

$$\sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\hat{\Omega}_{\hat{k}} - \Omega\|_{\mathrm{op}}^2 \le Cn^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n}.$$

$$\sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\hat{\Omega}_{\hat{k}} - \Omega\|_{\mathrm{op}}^2 \le Cn^{-\frac{2\alpha}{2\alpha+1}} + C\frac{\log p}{n}.$$

An Extension to Nonparanormal distributions

- Instead of $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T \sim N(0, \Omega^{-1})$, one only observe its transformed variables, $\mathbf{Y} = (f_1(X_1), f_2(X_2), \ldots, f_p(X_p))^T$, where $\{f_i\}_{i=1}^p$ are some unknown strictly increasing functions.
- Goal: Estimate the inverse of correlation matrix.
- Procedures: local sample covariance replaced by rank-based correlation matrix (Kendall's tau and Spearman's rho.)
- Analysis: Variance terms can be controlled by concentration inequalities of rank-based correlation matrices (e.g., [Mitra and Zhang, 2014]).

Minimax risk under Frobenius norm

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta, M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \qquad \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta, M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2$$

Minimax risk under Frobenius norm

Theorem 3 (Minimax risks under Frobenius norm)

The minimax risk of the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ satisfies

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \asymp \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \asymp n^{-\frac{2\alpha+1}{2\alpha+2}}$$

this rate can be achieved by the estimator defined as following.

Minimax risk under Frobenius norm

Theorem 3 (Minimax risks under Frobenius norm)

The minimax risk of the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ satisfies

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \asymp \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \asymp n^{-\frac{2\alpha+1}{2\alpha+2}}$$

this rate can be achieved by the estimator defined as following.

Remark: Since $\mathcal{Q}_\alpha(\eta, \alpha M) \subset \mathcal{P}_\alpha(\eta, M)$, it suffices to show the upper bound for $\mathcal{Q}_\alpha(\eta, M)$ and the matching lower bound for $\mathcal{Q}_\alpha(\eta, M)$

Estimation procedure: regression-based estimator

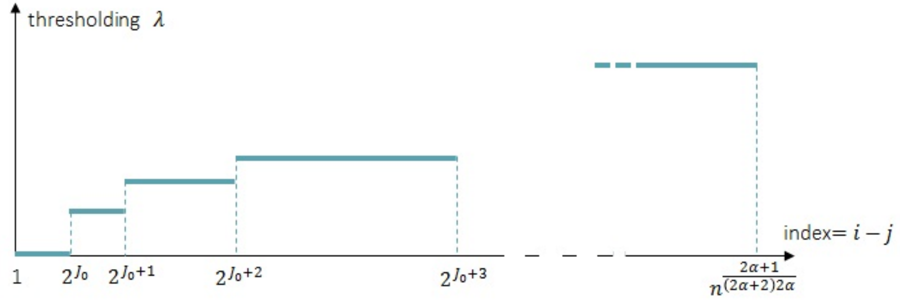
$$\tilde{\Omega}_k^{\mathrm{F}} = (I - \tilde{A})^T \tilde{D}^{-1} (I - \tilde{A}).$$

- Step 1: First regress X_i against $\mathbf{X}_{i-k_1:i-1} = (X_{i-k_1}, \ldots, X_{i-1})^T$ with a slightly larger bandwidth $k_1 = \lceil n^{\frac{2\alpha+1}{(2\alpha+2)2\alpha}} \rceil$ to obtain $\hat{\mathbf{a}}_i$;
- Step 2: Apply the block-thresholding rule

$$\hat{a}_{ij}^* = \hat{a}_{ij} \mathbf{1} \left(|\hat{a}_{ij}| > \lambda_j \right), i - k_1 \leq j \leq i - 1, \tag{1}$$

  where $\lambda_j = (\lceil \log_2^{i-j} - \log_2^{k_0} \rceil R)^{1/2}$ with $R = \eta \|(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\mathrm{op}}$ and $k_0 = n^{\frac{1}{2\alpha+2}}$. Set $\tilde{A}$ by arranging $\hat{a}_{ij}^*$.
- Step 3: Estimate each d_i using sample variance of empirical residuals $\tilde{d}_i$ of the ith regression above. Set $\tilde{D} = diag(\tilde{d}_i)$:

Estimation procedure: regression-based estimator



Remark: Motivated by wavelet analysis over Besov balls.

Remark: For the space $\mathcal{Q}_\alpha(\eta, M)$, a simpler banding estimation scheme is able to achieve the minimax rates.

Lower bound

$$\mathcal{P}' = \left\{\Omega(\theta) : \Omega(\theta) = (I_p - A(\theta))^T (I_p - A(\theta)), \theta = \{\theta(i)\}, \theta(i) \in \Theta\right\}.$$

$$A(\theta) = \begin{bmatrix} \boxed{\begin{matrix} 0_k & 0_k \\ n^{-\frac{1}{2}}\theta(1) & 0_k \end{matrix}} & 0_{2k} & \cdots & 0_{2k} \\ 0_{2k} & \boxed{\begin{matrix} 0_k & 0_k \\ n^{-\frac{1}{2}}\theta(2) & 0_k \end{matrix}} & \cdots & 0_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{2k} & 0_{2k} & \cdots & \boxed{\begin{matrix} 0_k & 0_k \\ n^{-\frac{1}{2}}\theta(\frac{p}{2k}) & 0_k \end{matrix}} \end{bmatrix}.$$

where $\Theta = \{0, 1\}^{k \times k}$. $k = n^{\frac{1}{2\alpha+2}}$.

Summary

- We establish the minimax rates of convergence for estimating precision matrices with bandable Cholesky factor ($\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$) under both Operator norm and Frobenius norm.
- A striking phenomenon: Unlike the results for bandable covariance matrix estimation, estimating $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ are fundamental different under operator norm.
- Novel rate optimal procedures: Local cropping estimator and regression-based estimator with block-thresholding rule.
- An adaptive procedure: Lepski's method.
- An extension to nonparanormal models.

Summary

Comparison of minimax rates of estimating bandble covariance matrices [Cai et al., 2010].

bandable Cholesky factors

	Operator norm	Frobenius norm
$\mathcal{P}_\alpha(\eta, M)$	$n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}$,	$n^{-\frac{2\alpha+1}{2\alpha+2}}$
$\mathcal{Q}_\alpha(\eta, M)$	$n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$,	$n^{-\frac{2\alpha+1}{2\alpha+2}}$

bandable covariance matrices

	Operator norm	Frobenius norm
$\mathcal{P}_\alpha(\eta, M)$	$n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$,	$n^{-\frac{2\alpha+1}{2\alpha+2}}$
$\mathcal{Q}_\alpha(\eta, M)$	$n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$,	$n^{-\frac{2\alpha+1}{2\alpha+2}}$

Bickel, P. J. and Levina, E. (2008a).
Covariance regularization by thresholding.
The Annals of Statistics, pages 2577–2604.

Bickel, P. J. and Levina, E. (2008b).
Regularized estimation of large covariance matrices.
The Annals of Statistics, pages 199–227.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010).
Optimal rates of convergence for covariance matrix estimation.
The Annals of Statistics, 38(4):2118–2144.

Cai, T. T. and Zhou, H. H. (2012).
Optimal rates of convergence for sparse covariance matrix estimation.
The Annals of Statistics, pages 2389–2420.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008).
First-order methods for sparse covariance selection.
SIAM Journal on Matrix Analysis and Applications, 30(1):56–66

Johnstone, I. M. (2001).
On the distribution of the largest eigenvalue in principal components analysis.
Annals of statistics, pages 295–327.

Lee, K. and Lee, J. (2017).
Estimating large precision matrices via modified cholesky decomposition.
arXiv preprint arXiv:1707.01143.

Meinshausen, N. and Bühlmann, P. (2006).
High-dimensional graphs and variable selection with the lasso.
The Annals of Statistics, pages 1436–1462.

Mitra, R. and Zhang, C.-H. (2014).
Multivariate analysis of nonparametric estimates of large correlation matrices.
arXiv preprint arXiv:1403.6195.

Ren, Z., Sun, T., Zhang, C.-H., Zhou, H. H., et al. (2015).

Asymptotic normality and optimalities in estimation of large gaussian graphical models.
The Annals of Statistics, 43(3):991–1026.