Inference and Optimalities in Estimation of Gaussian Graphical Model

Harrison H. Zhou
Department of Statistics
Yale University

Jointly with Zhao Ren, Tingni Sun and Cun-Hui Zhang

Outline

- **Introduction**
- **Main Results**
  - **Asymptotic Efficiency**
  - **Rate-optimal Estimation of Each Entry**
- **Applications**
  - **Adaptive Support Recovery**
  - **Estimation Under the Spectral Norm**
  - **Latent Variable Graphical Model**
- **Summary**

Introduction

Gaussian Graphical Model:

Let $G = (V, E)$ be a graph. $V = \{Z_1, \ldots, Z_p\}$ is the vertex set and E is the edge set representing conditional dependence relations between the variables.

Consider

$$Z = (Z_1, Z_2, \ldots, Z_p)^T \sim \mathcal{N}\left(0, \Omega^{-1}\right),$$

where $\Omega = (\omega_{ij})_{1 \le i,j \le p}$.

Question:

Are Z_i and Z_j conditionally independent given $Z_{\{i,j\}^c}$?

Conditional Independence

Property:

The conditional distribution of Z_A given Z_{A^c} is

$$Z_A | Z_{A^c} = \mathcal{N}\left(-\Omega_{A,A}^{-1}\Omega_{A,A^c}Z_{A^c}, \Omega_{A,A}^{-1}\right),$$

where $A \subset \{1, 2, \ldots, p\}$.

Example:

Let $A = \{1, 2\}$. The precision matrix of $(Z_1, Z_2)^T$ given $Z_{\{1,2\}^c}$ is

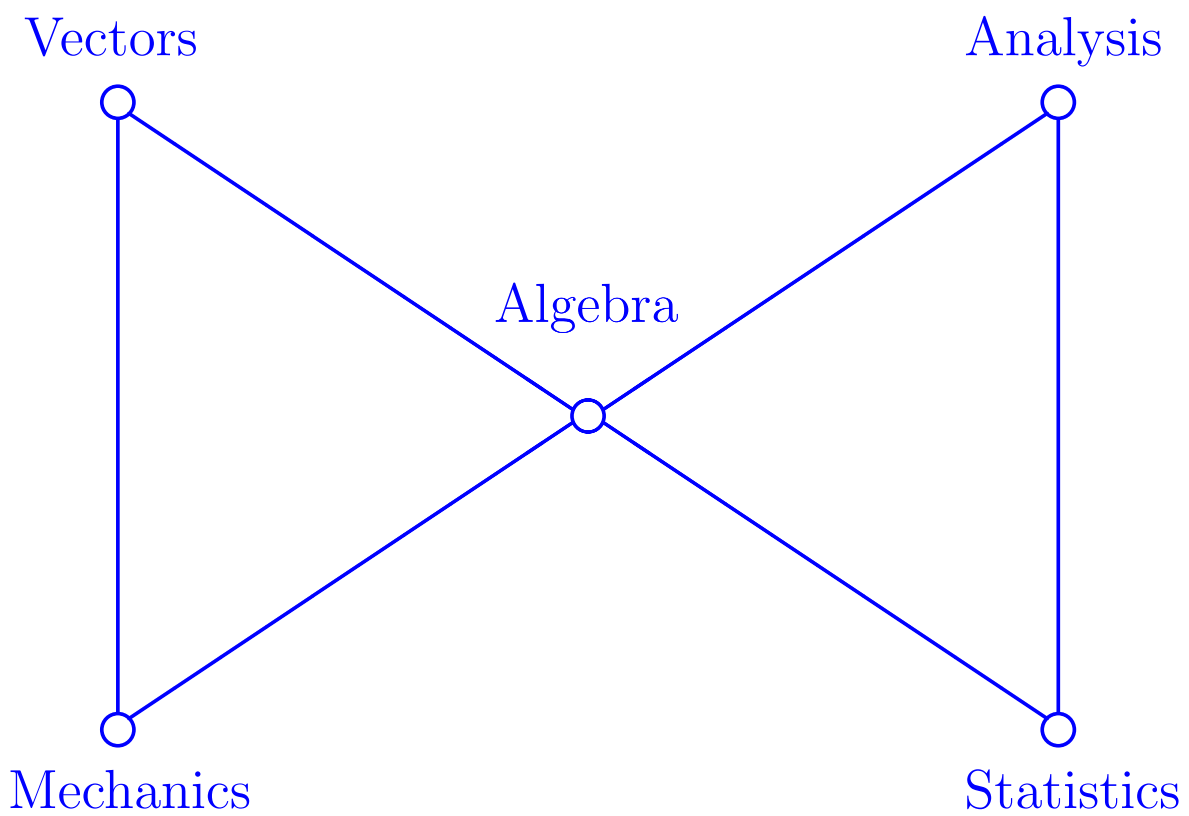
$$\Omega_{A,A} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}.$$

Hence

$$Z_1 \perp Z_2 | Z_{\{1,2\}^c} \Longleftrightarrow \omega_{12} = 0.$$

An Old Example

Whittaker (1990): Examination marks of 88 students in 5 different mathematical subjects, Analysis, Statistics, Mechanics, Vectors, Algebra.



Remark {Analysis, Stats} $\perp$ {Mech, Vectors} | Algebra.

What to do when p is very large?

Assumptions

Consider a class of sparse precision matrices $\mathcal{G}_0(M, k_{n,p})$:

- For $\Omega = (\omega_{ij})_{1 \le i,j \le p}$,

$$\max_{1 \le j \le p} \sum_{i \ne j} 1\left\{\omega_{ij} \ne 0\right\} \le k_{n,p},$$

  where $1\left\{\cdot\right\}$ is the indicator function.

- In addition, we assume $1/M \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le M$, for some constant $M > 1$.

GLASSO

Penalized Estimation:

$$\hat{\Omega}_{\text{Glasso}} := \arg\min_{\Omega \succ 0} \{ \langle \Omega, \Sigma_n \rangle - \log\det(\Omega) + \lambda_n |\Omega|_{1,\text{off}} \}$$

where Σ_n is the sample covariance of sample size n, and $|\Omega|_{1,\text{off}} = \sum_{i \neq j} |\omega_{ij}|$ is the vector ℓ_1 norm of off-diagonal elements.

GLASSO

Ravikumar, Wainwright, Raskutti and Yu (2011).

Assumptions:

- **Irrepresentable Condition:** There exists some $\alpha \in (0, 1]$ such that

$$\|\Gamma_{S^c S}(\Gamma_{SS})^{-1}\|_\infty \leq 1 - \alpha,$$

  where $\Gamma = \Omega_0^{-1} \otimes \Omega_0^{-1}$ and $S = \text{supp}(\Omega_0)$. $\|A\|_\infty$ is the maximum row absolute sum of A.

- For **support recovery**, the nonzero entry needs to be at least at an order of

$$\|(\Gamma_{SS})^{-1}\|_\infty \left(\frac{\log p}{n}\right)^{1/2},$$

  under the assumption that $k_{n,p} = o\left(\sqrt{n}/\log p\right)$.

Remarks:

- Meinshausen and Buhlmann (2006).
- Cai, Liu and Luo (2010) and Cai, Liu and Z. (2012, sumitted).

Main Results

Basic Property:

Let $A = \{1, 2\}$. The conditional distribution of Z_A given Z_{A^c} is

$$Z_A | Z_{A^c} = \mathcal{N}\left(-\Omega_{A,A}^{-1}\Omega_{A,A^c}Z_{A^c}, \Omega_{A,A}^{-1}\right),$$

where

$$\Omega_{A,A} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix},$$

and Ω_{A,A^c} is the first two rows of the precision matrix Ω.

Remark:

More generally we may consider $A = \{i, j\}$ or a finite subset.

Methodology

Let $X^{(i)} \stackrel{\text{i.i.d.}}{\sim} N_p(0, \Sigma)$, $i = 1, 2, \ldots, n$.

Let $\mathbf{X}$ be the data matrix of size n by p.

Let $\mathbf{X}_A$ be the columns indexed by $A = \{1, 2\}$ of size n by 2.

Regression

$$\mathbf{X}_A = \mathbf{X}_{A^c} \beta + \epsilon_A,$$

where $\beta^{\mathbf{T}} = -\Omega_{A,A}^{-1} \Omega_{A,A^c}$, and ϵ_A is an n by 2 matrix.

Methodology

Since

$$Z_A | Z_{A^c} = \mathcal{N}\left(-\Omega_{A,A}^{-1}\Omega_{A,A^c}Z_{A^c}, \Omega_{A,A}^{-1}\right),$$

we have

$$\mathbb{E}\epsilon_A^T\epsilon_A/n = \Omega_{A,A}^{-1}.$$

Efficiency

If you know β, an asymptotically efficient estimator is

$$\hat{\Omega}_{A,A} = \left(\epsilon_A^T\epsilon_A/n\right)^{-1}.$$

Methodology

Penalized Estimation

$$\left\{\hat{\beta}_m, \hat{\theta}_{mm}^{1/2}\right\} = \arg\min_{b \in \mathbb{R}^{p-2}, \sigma \in \mathbb{R}} \left\{ \frac{\|\mathbf{X}_m - \mathbf{X}_{A^c} b\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_{k \in A^c} \frac{\|\mathbf{X}_k\|}{\sqrt{n}} |b_k| \right\},$$

where $\lambda = \sqrt{\frac{2 \log p}{n}}$.

Residuals

$$\hat{\epsilon}_A = \mathbf{X}_A - \mathbf{X}_{A^c} \hat{\beta}.$$

Estimation

$$\hat{\Omega}_{A,A} = \left(\hat{\epsilon}_A^T \hat{\epsilon}_A / n\right)^{-1}.$$

Assumptions

Consider a class of sparse precision matrices $\mathcal{G}_0(M, k_{n,p})$:

- For $\Omega = (\omega_{ij})_{1 \le i,j \le p}$,

$$\max_{1 \le j \le p} \sum_{i \ne j} 1\left\{\omega_{ij} \ne 0\right\} \le k_{n,p},$$

  where $1\left\{\cdot\right\}$ is the indicator function.

- In addition, we assume $1/M \le \lambda_{\min}\left(\Omega\right) \le \lambda_{\max}\left(\Omega\right) \le M$, for some constant $M > 1$.

Remark

We actually consider a slightly more general definition of sparseness

$$\max_j \Sigma_{i \ne j} \min\left\{1, \left|\omega_{ij}\right| / \sqrt{\frac{2 \log p}{n}}\right\} \le k_{n,p}.$$

Asymptotic Efficiency

Theorem

Under the assumption that $k_{n,p} = o\left(\sqrt{n}/\log p\right)$ we have

$$\sqrt{nF_{ij}}\left(\hat{\omega}_{ij} - \omega_{ij}\right) \xrightarrow{D} \mathcal{N}\left(0,1\right),$$

where $F_{ij}^{-1} = \omega_{ii}\omega_{jj} + \omega_{ij}^2$.

Remark

We have a moderate deviation tail bound for $\hat{\omega}_{ij}$.

Optimality

Theorem

Under the assumption that $k_{n,p} = O\left(n/\log p\right)$ we have

$$\inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{E}\left|\hat{\omega}_{ij} - \omega_{ij}\right| \asymp \max\left\{k_{n,p}\frac{\log p}{n}, \sqrt{\frac{1}{n}}\right\},$$

under the assumption that $p \geq k_{n,p}^{\nu}$ with some $\nu > 2$.

Remark

- The upper bound is attained by our procedure.
- A necessary condition for estimating ω_{ij} consistently is $k_{n,p} = o\left(n/\log p\right)$.
- A necessary condition to obtain a parametric rate is, $k_{n,p}\frac{\log p}{n} = O(\sqrt{1/n})$, i.e., $k_{n,p} = O\left(\sqrt{n}/\log p\right)$.

Applications

Adaptive Support Recovery

Procedure

Let $\hat{\Omega}_{thr} = (\hat{\omega}_{ij}^{thr})_{p\times p}$ with

$$\hat{\omega}_{ij}^{thr} = \hat{\omega}_{ij} 1 \left\{ |\hat{\omega}_{ij}| \geq \delta \sqrt{\frac{\left(\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2\right) \log p}{n}} \right\}, \ \delta > 2$$

Assumption

$$|\omega_{ij}| \geq 2\delta \sqrt{\frac{\left(\omega_{ii}\omega_{jj} + \omega_{ij}^2\right) \log p}{n}}, \ \delta > 2, for \ \omega_{ij} \neq 0$$

Theorem

Let $\mathcal{S}(\boldsymbol{\Omega}) = \{sgn(\omega_{ij}), \ \ 1 \leq i, j \leq p\}$. We have

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{S}(\hat{\Omega}_{thr}) = \mathcal{S}(\Omega)\right) = 1,$$

provided that $k_{n,p} = o\left(\sqrt{n}/\log p\right)$.

Estimation Under the Spectral Norm

Procedure

Let $\hat{\Omega}_{thr} = (\hat{\omega}_{ij}^{thr})_{p\times p}$ with

$$\hat{\omega}_{ij}^{thr} = \hat{\omega}_{ij} 1 \left\{ |\hat{\omega}_{ij}| \geq \delta \sqrt{\frac{\left(\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2\right) \log p}{n}} \right\}, \ \delta > 2.$$

Theorem

The estimator $\hat{\Omega}_{thr}$ satisfied

$$\left\| \hat{\Omega}_{thr} - \Omega \right\|_{spectral}^2 = O_P\left(k_{n,p}^2 \frac{\log p}{n} \right),$$

uniformly over $\Omega \in \mathcal{G}_0(M, k_{n,p})$, provided that $k_{n,p} = o\left(\sqrt{n}/\log p\right)$.

Remark

Cai, Liu and Z. (2012) showed the rate is optimal.

Latent Variable Graphical Model

- Let $G = (V, E)$ be a graph. $V = \{Z_1, \ldots, Z_{p+r}\}$ is the vertex set and E is the edge set. Assume that the graph is sparse.
- But we only observe $\mathbf{X} = (Z_1, \ldots, Z_p)$ is multivariate normal with a precision matrix Ω.
- It can be shown that Ω can be decomposed as the sum of a sparse matrix and a rank r matrix by the Schur complement.

Question:

How to estimate Ω based on $\{X_i\}$, when $\Omega = (\omega_{ij})$ can be decomposed as the sum of a sparse matrix S and a rank r matrix L, i.e., $\Omega = S + L$?

Sparse + Low Rank

- **Sparse**

$$\mathcal{G}(k_{n,p}) = \left\{ S = (s_{ij}) : S \succ 0, \ \max_{1 \le i \le p} \sum_{j=1}^{p} \mathbf{1}\{s_{ij} \neq 0\} \le k_{n,p} \right\}$$

- **Low Rank**

$$L = \sum_{i=1}^{r} \lambda_i u_i u_i^T,$$

where there exists a universal constant c_0 such that $\|u_i\|_\infty \le \sqrt{\frac{c_0}{p}}$ for all i, and λ_i is bounded for all i by M. See Candès, Li, Ma, and Wright (2009).

- In addition, we assume $1/M \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le M$, for some constant $M > 1$.

Penalized Maximum Likelihood

Chandrasekaran, Parrilo and Willsky (2012, AoS)

Algorithm:

$$\hat{\Omega}_{\text{Glasso}} := \arg\min_{\Omega \succ 0} \{\langle \Omega, \Sigma_n \rangle - \log\det(\Omega) + \lambda_n |S|_1 + \gamma_n \|L\|_{nuclear}\}$$

Notations:

Minimum magnitude of nonzero entries of S by θ, i.e.,
$\theta = \min_{i,j} |s_{ij}| \mathbf{1}\{s_{ij} \neq 0\}$.

Minimum nonzero singular values of L by σ, i.e., $\sigma = \min_{1 \leq i \leq r} \lambda_i$.

Chandrasekaran, Parrilo and Willsky (2012, AoS)

To estimate the support and rank **consistently**, assuming that the authors can pick the tuning parameters "wisely" (as they wish), they still require:

- $\theta \gtrsim \sqrt{p/n}$
- $\sigma \gtrsim k_{n,p}^3 \sqrt{p/n}$

in addition to the strong irrepresentability condition and assumptions on the Fisher information matrix, and possibly other assumptions . . . .

Remark

Ren and Z. (2012) showed conditions can be significantly improved.

Optimality

Theorem

Assume that $p \geq \sqrt{n}$. We have

$$|\hat{\Omega} - \Omega|_\infty = O_P\left(\sqrt{\frac{\log p}{n}}\right),$$

provided that $k_{n,p} = o(\sqrt{n/\log p})$.

Remark

- We can do adaptive support recovery similar to the sparse case. Improve the order of θ from $\sqrt{p/n}$ to $\sqrt{\log(p)/n}$ (optimal).
- To estimate the rank consistently we improve the order of σ from $k_{n,p}^3\sqrt{p/n}$ to $\sqrt{p/n}$ (optimal).

Summary

- A methodology to do inference.
- A necessary sparseness condition for inference.
- Applications to adaptive support recovery, optimal estimation under the spectral norm and latent variable graphical model.