High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence

Maziar Sanjabi

University of Minnesota

April 10, 2014

- P. Ravikumar, M. J. Wainwright, G. Raskutti and B. Yu. *"High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence."*

Outline

Motivation

Outline

Outline

Outline

Outline

Gaussian Graphical Model

- $(X_1, \cdots, X_p)$ zero mean Gaussian RV with covariance Σ^*.
- $\Theta^* = (\Sigma^*)^{-1}$

$$f(x_1, \cdots, x_p; \Theta^*) = \frac{1}{\sqrt{(2\pi)^p \det((\Theta^*)^{-1})}} \exp\left\{ -\frac{1}{2}\mathbf{x}^T \Theta^* \mathbf{x} \right\}.$$

- Given n sample vectors $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}$, ML estimator of Θ^*

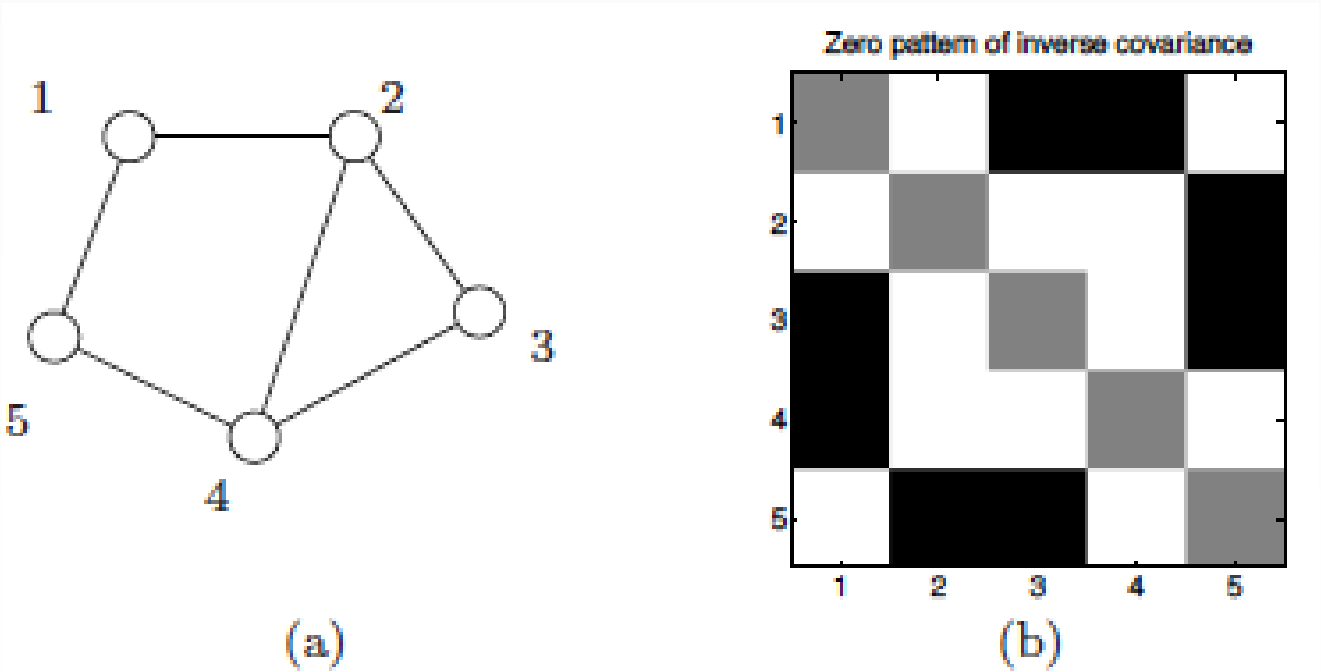
$$\widehat{\Theta} = \arg\max_{\Theta \succ 0} \log\det(\Theta) - \langle \Theta, \widehat{\Sigma}^n \rangle,$$

where $\widehat{\Sigma}^n = \frac{1}{n}\sum_{k=1}^n \mathbf{x}^{(k)}(\mathbf{x}^k)^T$ is sample covariance.

Gaussian Graphical Model (Con'd)

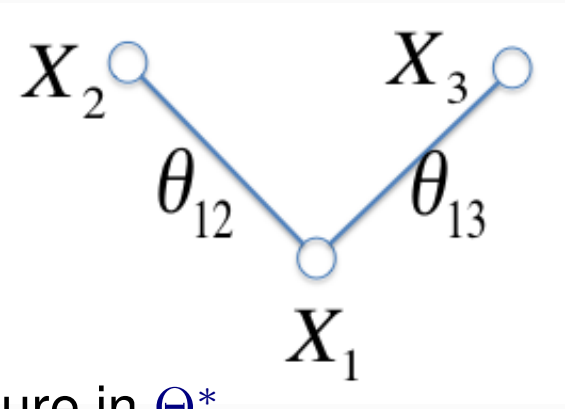
- Model the dependencies based on a graph.
- Consider graph $G = (V, E)$, each node in V corresponds to a variable.
- Conditional independence

$$(i, j) \notin E \Rightarrow X_i \perp X_j | X_{-ij} \Rightarrow \Theta^*_{ij} = 0.$$



Gaussian Graphical Model (Con'd)

$$\begin{pmatrix} \Theta_{11} & \Theta_{12} & \Theta_{13} \\ \Theta_{21} & \Theta_{22} & 0 \\ \Theta_{13} & 0 & \Theta_{33} \end{pmatrix}$$



- Few edges in E $\Rightarrow$ Sparse off-diagonal structure in Θ^*.
- Exploiting sparsity in solution

$$\widehat{\Theta} \;=\; \arg\min_{\Theta \succ 0} \langle \Theta, \widehat{\Sigma}^n \rangle - \log\det(\Theta) + \lambda_n \|\Theta\|_{1,o}, \qquad \text{(P)}$$

$\|\Theta\|_{1,o} = \sum_{i \neq j} |\Theta_{ij}|$

- Set $E(\Theta^*) = \{(i,j) \;\mid\; i \neq j,\; \Theta^*_{ij} \neq 0\}$. $s = |E(\Theta^*)|$.
- Maximum degree $d = \max_i \,|\, \{j \in V \mid \Theta^*_{ij} \neq 0\} \,|$

Tail Conditions

- Closeness of sample covariance to true covariance.

Tail Conditions

For any small enough δ, $\forall\ i,\ j$

$$\mathbb{P}[|\widehat{\Sigma}^n_{i,j} - \Sigma^*_{i,j}| > \delta] \leq \frac{1}{f(n,\delta)}.$$

- Exponential: $f(n,\delta) = \exp(cn\delta^a)$, $a > 0$; Sub-Gaussian: $f(n,\delta) = \exp(cn\delta^2)$
- Polynomial: $f(n,\delta) = cn^m\delta^{2m}$, $m \in \mathbb{N}$: RV with bounded $4m$-th moment.

- How many samples needed s.t. $\mathbb{P}[\cdot] \leq 1/r$: $\bar{n}_f(\delta, r)$.
- What is maximum accuracy with fixed probability $1/r$: $\bar{\delta}_f(n, r)$.

$$n > \bar{n}_f(\delta, r) \ \Rightarrow\ \bar{\delta}_f(n, r) < \delta$$

Conditions on Covariance and Hessian

Hessian of the objective at Θ^* is

$$\Gamma^* = \Theta^{*-1} \otimes \Theta^{*-1}.$$

- Gaussian: Γ^* Fisher Information
- $S(\Theta^*) = E(\Theta^*) \cup \{(1,1), \cdots, (p,p)\}$, $|S| = s + p$.
- Define

$$K_{\Sigma^*} = |||\Sigma^*|||_\infty = \max_i \sum_j |\Sigma^*_{ij}|$$

- $\Gamma^*_{SS} = [\Theta^{*-1} \otimes \Theta^{*-1}]_{SS}$

$$K_{\Gamma^*} = \left|\left|\left|\Gamma^{*-1}_{SS}\right|\right|\right|_\infty.$$

Mutual Incoherence or Irrepresentability Condition

Assumption 1

There exists $\alpha \in (0, 1]$

$$\left\| \left| \Gamma^*_{S^c S} (\Gamma^*_{SS})^{-1} \right| \right\|_\infty \leq (1 - \alpha).$$

- Limits the influence of non-edge terms (in S^c) on edge terms.
- Similar to incoherence condition for Lasso; but on edge variables.

First Result

- $\tau > 2$, higher $\tau \Rightarrow$ higher probability guarantee but more samples

Theorem 1

- $r = p^\tau$;
- $\delta = 1/C_1 d$ (C_1 depends on α and K)
- $\lambda_n = \frac{8}{\alpha}\bar{\delta}_f(n, r)$

If $n > \bar{n}_f(\delta, r)$, with probability $1 - 1/p^{\tau-2} \to 1$

$$\boxed{\|\widehat{\Theta} - \Theta^*\|_\infty \leq C_2 \bar{\delta}_f(n, p^\tau)}$$

Moreover, $E(\widehat{\Theta}) \subset E(\Theta^*)$ and includes edges s.t.

$$\boxed{|\Theta^*_{ij}| > C_2 \bar{\delta}_f(n, r)}$$

- Easily generalize to $\|\cdot\|_F$ and $\|\cdot\|_2$.

Sub-Gaussian

For sub-Gaussian

$$\bar{n}_f(\delta, r) = \frac{\log r}{c\delta^2}$$

$$\bar{\delta}_f(n, r) = \sqrt{\frac{\log r}{cn}}$$

Proposition 1

For sub-Gaussian RVs if $n > C_1 d^2 \tau \log p$, then with probability $1 - 1/p^{\tau-2}$

$$\boxed{\|\widehat{\Theta} - \Theta^*\|_\infty \leq C_2 \sqrt{\frac{\tau \log p}{n}}}$$

and any edge is recovered if

$$\boxed{|\Theta^*_{ij}| > C_2 \sqrt{\frac{\tau \log p}{n}}}$$

$n = \Omega(d^2 \log p)$

Polynomial Tail

For polynomial tail

$$\bar{n}_f(\delta, r) = \frac{r^{1/m}}{c\delta^2}$$

$$\bar{\delta}_f(n, r) = \sqrt{\frac{r^{1/m}}{n}}$$

Proposition 2

For RVs with polynomial tail if $n > C_1 d^2 p^{\tau/m}$, then with probability $1 - 1/p^{\tau-2}$

$$\boxed{\|\widehat{\Theta} - \Theta^*\|_\infty \leq C_2 \sqrt{\frac{p^{\tau/m}}{n}}}$$

and any edge is recovered if

$$\boxed{|\Theta^*_{ij}| > C_2 \sqrt{\frac{p^{\tau/m}}{n}}}$$

$n = \Omega(d^2 p^{\tau/m})$

Model Selection Consistency

- Define $\theta_{\min} := \min_{(i,j)\in E(\Theta^*)} |\Theta^*_{ij}|$ and

$$\mathcal{M}(\widehat{\Theta}; \Theta^*) := \{sign(\widehat{\Theta}_{ij}) = sgn(\Theta^*_{ij}) \; \forall \; (i,j) \in E(\Theta^*)\}.$$

Theorem 2

With similar conditions if $\delta = \frac{1}{\max(C_0 \theta_{\min}^{-1}, C_1 d)}$ and $n > \bar{n}_f(\delta, p^\tau)$ then

$$\mathbb{P}[\mathcal{M}(\widehat{\Theta}; \Theta^*)] \geq 1 - 1/p^{\tau-2} \to 1.$$

Proposition 3

- Exponential tail

$$n = \Omega((d^2 + \theta_{\min}^{-2})\tau \log p)$$

- Polynomial tail

$$n = \Omega((d^2 + \theta_{\min}^{-2})p^{\tau/m})$$

Primal-dual witness approach

1. Construct a solution $\widetilde{\Theta}$: complying with $E(\Theta^*)$.
2. Construct dual $\widetilde{Z} \overset{?}{\in} \partial\|\widetilde{\Theta}\|_{1,o}$ s.t. $(\widetilde{\Theta}, \widetilde{Z})$ satisfies optimality.
3. Verify $\widetilde{Z} \in \partial\|\widetilde{\Theta}\|_{1,o}$, if $\|\widetilde{\Theta} - \Theta^*\|$ and $\|\widehat{\Sigma} - \Sigma^*\|$ small enough.
4. (P) has unique solution $\Rightarrow \widehat{\Theta} = \widetilde{\Theta}$.
5. Bound $\|\widetilde{\Theta} - \Theta^*\|$ and $\|\widehat{\Sigma} - \Sigma^*\|$ with high probability.

Step 4 (Uniqueness)

$$\widehat{\Theta} \;=\; \arg\min_{\Theta \succ 0} \langle \Theta, \widehat{\Sigma}^n \rangle - \log\det(\Theta) + \lambda_n \|\Theta\|_{1,o}, \tag{P}$$

Lemma 1

For any $\lambda_n > 0$, if $\widehat{\Sigma}_{ii} > 0$, $\Rightarrow$ (P) has unique solution satisfying

$$\widehat{\Sigma} - \widehat{\Theta}^{-1} + \lambda_n \widehat{Z} = 0, \tag{1}$$

$\widehat{Z} \in \partial \|\widehat{\Theta}\|_{1,o}$.

- strictly convex, easy to prove coercive $\Rightarrow$ unique solution.

Constructing Primal-Dual Solution (Steps 1 and 2)

- Knowing S

$$\widetilde{\Theta} := \arg \min_{\Theta_{S^c}=0} \{\langle \Theta, \widehat{\Sigma} \rangle - \log \det(\Theta) + \lambda_n \|\Theta\|_{1,o}\}.$$

- Choose $\widetilde{Z} \in \partial \|\widetilde{\Theta}\|_{1,o}$ ($\widetilde{Z}_{ij} = sign(\widetilde{\Theta}_{ij})$ if $i \neq j$ and $\Theta_{ij} \neq 0$).

$$\widehat{\Sigma}_{ij} - [\widetilde{\Theta}^{-1}]_{ij} + \lambda_n \widetilde{Z}_{ij} = 0, \ (i,j) \in S.$$

- Edit $\widetilde{Z}_{ij}$ for $(i,j) \in S^c$ to satisfy optimality condition of (P)

$$\widetilde{Z}_{ij} := \frac{1}{\lambda_n} \{-\widehat{\Sigma}_{ij} + [\widetilde{\Theta}^{-1}]_{ij}\}. \tag{2}$$

- Verify strict dual feasibility for step 3

$$|\widetilde{Z}_{ij}| < 1, \ (i,j) \in S^c. \tag{3}$$

Step 3

Define $\Delta = \widetilde{\Theta} - \Theta^*$, $W = \widehat{\Sigma} - \Sigma^*$.

$$R(\Delta) = \widetilde{\Theta}^{-1} - \Theta^{*-1} + \Theta^{*-1}\Delta\Theta^{*-1}. \tag{4}$$

Lemma 2

If $\max\{\|W\|_\infty, \|R(\Delta)\|_\infty\} \leq \frac{\alpha\lambda_n}{8}$, then $\widetilde{Z}$ is a valid sub-gradient $\Rightarrow \widetilde{\Theta} = \widehat{\Theta}$.

Proof ingridients:

- Rewrite optimality as

$$\Theta^{*-1}\Delta\Theta^{*-1} + W - R + \lambda_n\widetilde{Z} = 0. \tag{5}$$

- Vectorize: $\Theta^{*-1}\Delta\Theta^{*-1} = \Gamma^*\bar{\Delta}$. Note $\bar{\Delta}_{S^c} = 0$.

$$\Gamma^*_{SS}\bar{\Delta}_S + \bar{W}_S - \bar{R}_S + \lambda_n\bar{\widetilde{Z}}_S = 0.$$

$$\Gamma^*_{S^cS}\bar{\Delta}_S + \bar{W}_{S^c} - \bar{R}_{S^c} + \lambda_n\bar{\widetilde{Z}}_{S^c} = 0. \tag{6}$$

- Use Assumption 1.

Step 3 (Con'd)

Relating $\|R(\Delta)\|_\infty$ to $\|\Delta\|_\infty$

Lemma 3

If $\|\Delta\|_\infty \leq \frac{1}{3K_{\Sigma^*}d} \Rightarrow \|R(\Delta)\|_\infty \leq \frac{3}{2}d\|\Delta\|_\infty{}^2 K_{\Sigma^*}^3$

- Combining Lemma 1, 2 and 3 $\Rightarrow \widetilde{\Theta} = \widehat{\Theta}$ if $\|\Delta\|_\infty$ and $\|W\|_\infty$ small enough.
- Step 5: Probabilistically bound $\|\Delta\|_\infty$ and $\|W\|_\infty$.
- High probability bounds for $\|W\|_\infty$: easy; tail conditions
- Bound $\|\Delta\|_\infty$ by $\|W\|_\infty$.

Step 5

Lemma 4

If

$$r = 2K_{\Gamma^*}(\|W\|_\infty + \lambda_n) \leq \frac{c}{d},$$

then

$$\|\Delta\|_\infty \leq r$$

Sketch of Proof:

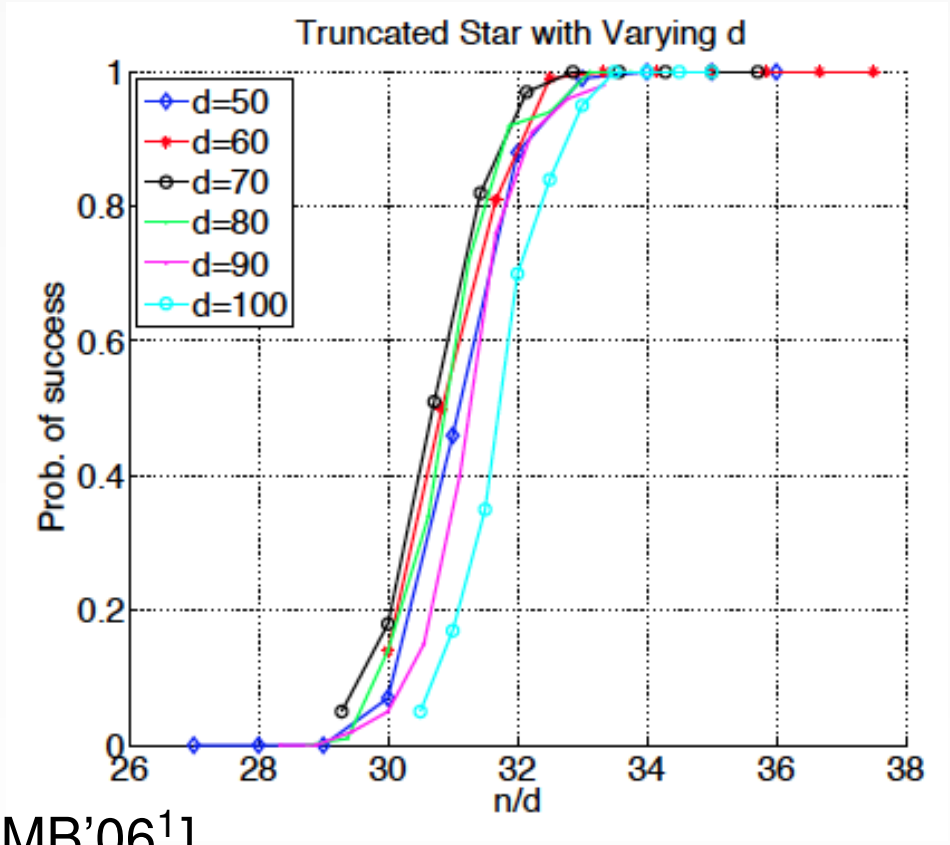
- Note $\bar{\Delta}_{S^c} = 0$.
- Construct continuous function $F : B(r) \to B(r)$ ($B(r) = \{\bar{\hat{\Delta}} | \bar{\hat{\Delta}}_{S^c} = 0, \|\bar{\hat{\Delta}}_S\|_\infty \leq r\}$) with only possible fixed point at Δ.
- Use Brouwer's fixed point: F has one fixed point.

Lemma 5

$$\mathbb{P}[\|W\|_\infty \geq \bar{\delta}_f(n, p^\tau)] \leq \frac{1}{p^{\tau-2}}$$

- *Proof*: tail condition + union bound.
- Combining Lemma 1-5 it is easy to prove Theorem 1.
- Similar proof for Theorem 2.

Insightful Experiments and Comparison



- $n = \Omega(d^2 \log p)$ conservative.
- $n = \Omega(d^\gamma \log p)$, $\gamma \in (1, 2)$.
- $n = \Omega(d \log p)$ achievable with neighborhood-based method [MB'06[1]]
  - Find regression vectors based using Lasso. Find neighbors based on support.
  - Combine these neighbor sets to form graph.

[1] "High-dimensional graphs and variable selection with the Lasso," N. Meinshausen and P. Buhlmann.

Thank You!