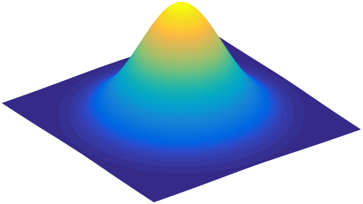
Gaussian Graphical Models and Graphical Lasso

Yuxin Chen

Princeton University, Fall 2020

Multivariate Gaussians

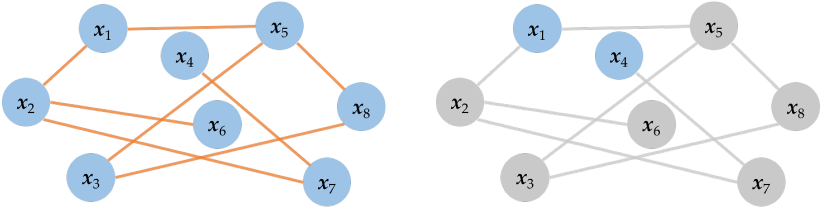


Consider a random vector $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with probability density

$$
\begin{aligned}
f(\boldsymbol{x}) \;&=\; \frac{1}{(2\pi)^{p/2} \det\left(\boldsymbol{\Sigma}\right)^{1/2}} \exp\left\{-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right\} \\
&\propto\; \det\left(\boldsymbol{\Theta}\right)^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Theta}\boldsymbol{x}\right\}
\end{aligned}
$$

where $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] \succ \mathbf{0}$ is the covariance matrix, and $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is the inverse covariance matrix or precision matrix

Undirected graphical models



$$x_1 \perp\!\!\!\perp x_4 \mid \{x_2, x_3, x_5, x_6, x_7, x_8\}$$

- Represent a collection of variables $\boldsymbol{x} = [x_1, \cdots, x_p]^\top$ by a vertex set $\mathcal{V} = \{1, \cdots, p\}$
- Encode conditional independence by a set $\mathcal{E}$ of edges
  - For any pair of vertices u and v,

$$(u, v) \notin \mathcal{E} \quad \Longleftrightarrow \quad x_u \perp\!\!\!\perp x_v \mid \boldsymbol{x}_{\mathcal{V} \setminus \{u,v\}}$$

Gaussian graphical models

Fact 11.1

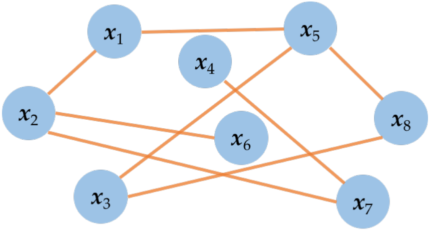
(Homework) Consider a Gaussian vector $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$*. For any* u *and* v*,*

$$x_u \perp\!\!\!\perp x_v \mid \boldsymbol{x}_{\mathcal{V}\backslash\{u,v\}}$$

iff $\Theta_{u,v} = 0$*, where* $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$

conditional independence $\Longleftrightarrow$ sparsity

Gaussian graphical models



$$\underbrace{\begin{bmatrix} * & * & 0 & 0 & * & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & * & * & 0 \\ 0 & 0 & * & 0 & * & 0 & 0 & * \\ 0 & 0 & 0 & * & 0 & 0 & * & 0 \\ * & 0 & * & 0 & * & 0 & 0 & * \\ 0 & * & 0 & 0 & 0 & * & 0 & 0 \\ 0 & * & 0 & * & 0 & 0 & * & 0 \\ 0 & 0 & * & 0 & * & 0 & 0 & * \end{bmatrix}}_{\boldsymbol{\Theta}}$$

Likelihoods for Gaussian models

Draw n i.i.d. samples $\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, then the log-likelihood (up to additive constant) is

$$\begin{aligned} \ell(\boldsymbol{\Theta}) &= \frac{1}{n} \sum_{i=1}^{n} \log f(\boldsymbol{x}^{(i)}) = \frac{1}{2} \log\det(\boldsymbol{\Theta}) - \frac{1}{2n} \sum_{i=1}^{n} \boldsymbol{x}^{(i)\top} \boldsymbol{\Theta} \boldsymbol{x}^{(i)} \\ &= \frac{1}{2} \log\det(\boldsymbol{\Theta}) - \frac{1}{2} \langle \boldsymbol{S}, \boldsymbol{\Theta} \rangle , \end{aligned}$$

where $\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)\top}$: sample covariance; $\langle \boldsymbol{S}, \boldsymbol{\Theta} \rangle = \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Theta})$

Maximum likelihood estimation

$$\text{maximize}_{\boldsymbol{\Theta} \succeq \mathbf{0}} \quad \log\det(\boldsymbol{\Theta}) - \langle \boldsymbol{S}, \boldsymbol{\Theta} \rangle$$

Challenge in high-dimensional regime

Classical theory says MLE coverges to the truth as sample size $n \to \infty$

Practically, we are often in the regime where the sample size n is small ($n < p$)

- In this regime, $\boldsymbol{S}$ is rank-deficient, and the MLE does not even exist (why?)

Graphical lasso (Friedman, Hastie, &Tibshirani '08)

In practice, many pairs of variables might be conditionally independent
$\Longleftrightarrow$ many missing links in the graphical model (sparsity)

Key idea: use ℓ_1 regularization to promote sparsity

$$\text{maximize}_{\boldsymbol{\Theta}\succeq\mathbf{0}} \quad \log\det(\boldsymbol{\Theta}) - \langle \boldsymbol{S}, \boldsymbol{\Theta}\rangle - \underbrace{\lambda\|\boldsymbol{\Theta}\|_1}_{\text{lasso penalty}}$$

- Convex program! (homework)

Graphical lasso (Friedman, Hastie, &Tibshirani '08)

$$\text{maximize}_{\boldsymbol{\Theta}\succeq \mathbf{0}} \quad \log\det(\boldsymbol{\Theta}) - \langle \boldsymbol{S}, \boldsymbol{\Theta}\rangle - \underbrace{\lambda\|\boldsymbol{\Theta}\|_1}_{\text{lasso penalty}}$$

- First-order optimality condition

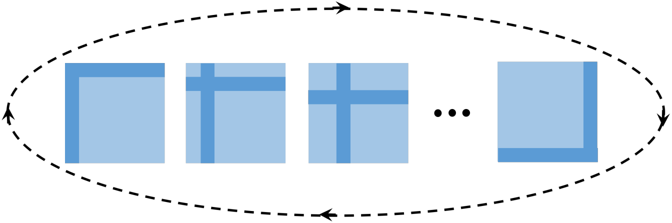
$$\mathbf{0} \in \boldsymbol{\Theta}^{-1} - \boldsymbol{S} - \lambda \underbrace{\partial\|\boldsymbol{\Theta}\|_1}_{\text{subdifferential}} \tag{11.1}$$

- For diagonal entries, one has $1 \in \partial|\Theta_{i,i}|$ (since $\Theta_{i,i} > 0$)

$$\Longrightarrow \quad (\boldsymbol{\Theta}^{-1})_{i,i} = S_{i,i} + \lambda, \quad 1 \le i \le p$$

(Optional) Blockwise coordinate descent

Idea: repeatedly cycle through all columns / rows and, in each step, optimize only a single column / row



Notation: use $\boldsymbol{W}$ to denote a working version of $\boldsymbol{\Theta}^{-1}$. Partition all matrices into 1 column / row vs. the rest

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}_{12}^\top & \theta_{22} \end{bmatrix} \quad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}^\top & s_{22} \end{bmatrix} \quad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}^\top & w_{22} \end{bmatrix}$$

(Optional) Blockwise coordinate descent

Blockwise step: suppose we fix all but the last row / column. It follows from (11.1) that

$$\mathbf{0} \in \boldsymbol{W}_{11}\boldsymbol{\beta} - \boldsymbol{s}_{12} - \lambda\partial\|\boldsymbol{\theta}_{12}\|_1 = \boldsymbol{W}_{11}\boldsymbol{\beta} - \boldsymbol{s}_{12} + \lambda\partial\|\boldsymbol{\beta}\|_1 \tag{11.2}$$

where $\boldsymbol{\beta} = -\boldsymbol{\theta}_{12}/\tilde{\theta}_{22}$ (since $\underbrace{\begin{bmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}_{12}^\top & \theta_{22} \end{bmatrix}^{-1} = \begin{bmatrix} * & -\frac{1}{\tilde{\theta}_{22}}\boldsymbol{\Theta}_{11}^{-1}\boldsymbol{\theta}_{12} \\ * & * \end{bmatrix}}_{\text{matrix inverse formula}}$) with

$\tilde{\theta}_{22} = \theta_{22} - \boldsymbol{\theta}_{12}^\top\boldsymbol{\Theta}_{11}^{-1}\boldsymbol{\theta}_{12} > 0$

This coincides with the optimality condition for

$$\text{minimize}_{\boldsymbol{\beta}} \quad \frac{1}{2}\left\|\boldsymbol{W}_{11}^{1/2}\boldsymbol{\beta} - \boldsymbol{W}_{11}^{-1/2}\boldsymbol{s}_{12}\right\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{11.3}$$

(Optional) Blockwise coordinate descent

Algorithm 11.1 Block coordinate descent for graphical lasso

Initialize $\boldsymbol{W} = \boldsymbol{S} + \lambda \boldsymbol{I}$ and fix its diagonals $\{w_{i,i}\}$.

Repeat until covergence:

for $t = 1, \cdots p$**:**

(i) Partition $\boldsymbol{W}$ (resp. $\boldsymbol{S}$) into 4 parts, where the upper-left part consists of all but the jth row / column

(ii) Solve

$$\text{minimize}_{\boldsymbol{\beta}} \quad \frac{1}{2} \| \boldsymbol{W}_{11}^{1/2} \boldsymbol{\beta} - \boldsymbol{W}_{11}^{-1/2} \boldsymbol{s}_{12} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_1$$

(iii) Update $\boldsymbol{w}_{12} = \boldsymbol{W}_{11} \boldsymbol{\beta}$

Set $\hat{\boldsymbol{\theta}}_{12} = -\hat{\theta}_{22} \boldsymbol{\beta}$ with $\hat{\theta}_{22} = 1/(w_{22} - \boldsymbol{w}_{12}^\top \boldsymbol{\beta})$

(Optional) Blockwise coordinate descent

The only remaining thing is to ensure $\boldsymbol{W} \succeq \mathbf{0}$. This is automatically satisfied:

Lemma 11.2 (Mazumder & Hastie '12)

If we start with $\boldsymbol{W} \succ \mathbf{0}$ satisfying $\|\boldsymbol{W} - \boldsymbol{S}\|_\infty \leq \lambda$, then every row / column update maintains positive definiteness of $\boldsymbol{W}$.

- If we start with $\boldsymbol{W}^{(0)} = \boldsymbol{S} + \lambda \boldsymbol{I}$, then $\boldsymbol{W}^{(t)}$ will always be positive definite

Reference

- "*Sparse inverse covariance estimation with the graphical lasso*," J. Friedman, T. Hastie, and R. Tibshirani, *Biostatistics*, 2008.
- "*The graphical lasso: new insights and alternatives*," R. Mazumder and T. Hastie, *Electronic journal of statistics*, 2012.
- "*Statistical learning with sparsity: the Lasso and generalizations*," T. Hastie, R. Tibshirani, and M. Wainwright, 2015.