ECE 18-898G: Special Topics in Signal Processing: Sparsity, Structure, and Inference

High-dimensional graphical models

Yuejie Chi

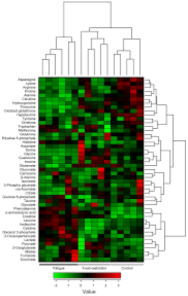
Department of Electrical and Computer Engineering

Carnegie Mellon University

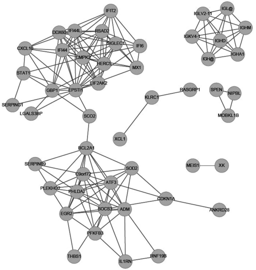
Spring 2018

Identifying Interactions in Data

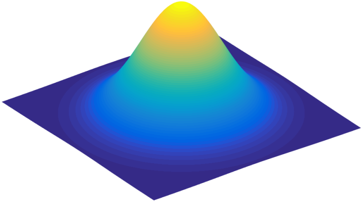
Given n data samples, $\boldsymbol{x}_i \sim \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \in \mathbb{R}^p$, how to identity interactions between x_i and x_j?



$\Longrightarrow$



Multivariate Gaussians



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

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

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

Likelihood function for Gaussian models

Draw n i.i.d. samples $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, then log-likelihood (up to additive constant) is

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

where $\boldsymbol{S} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ is sample covariance matrix (SCM).

Maximum likelihood estimation (MLE)

$$\widehat{\boldsymbol{\Theta}} = \mathsf{argmax}_{\boldsymbol{\Theta} \succeq \boldsymbol{0}} \quad \log\det(\boldsymbol{\Theta}) - \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Theta})$$

The sample-rich regime

Fact 9.1

If the SCM $\boldsymbol{S}$ is invertible, the MLE is given as

$$\widehat{\boldsymbol{\Theta}} = \boldsymbol{S}^{-1}.$$

When $n \gg p$, the SCM is invertible and classical theory says MLE converges to the truth as sample size $n \to \infty$ (consistency).

High-dimensional / sample-starved regime

Practically, we are often in the regime where sample size n is small, with $n < p$. Why?

- Our assumption may only hold for a small window of data collection;
- Our ability may only allow us to collect a few samples;
- The number of features/variables we care is much higher.

In this regime, $\boldsymbol{S}$ is rank-deficient, and MLE does not even exist.

High-dimensional / sample-starved regime

Practically, we are often in the regime where sample size n is small, with $n < p$. Why?

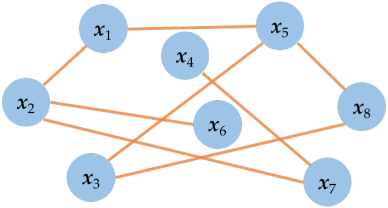
- Our assumption may only hold for a small window of data collection;
- Our ability may only allow us to collect a few samples;
- The number of features/variables we care is much higher.

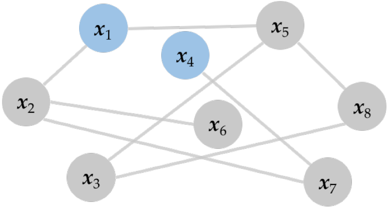
In this regime, $\boldsymbol{S}$ is rank-deficient, and MLE does not even exist.

Strategy: impose low-dimensional structures.

Gaussian Graphical Model with Sparsity

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

Lemma 9.2

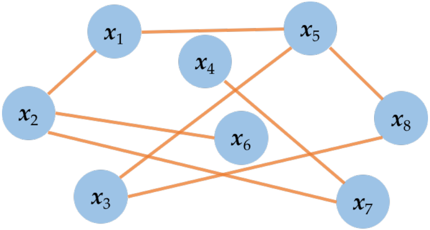
Consider a Gaussian vector $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. *For any* u *and* v,

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

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

Many pairs of variables are conditionally independent
$\Longleftrightarrow$ many missing links in the graphical model (sparsity)

Gaussian graphical models



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

Inverse covariance matrix $\boldsymbol{\Theta}$ is often (approximately) sparse

Sparse inverse covariance estimation

Problem definition: Given n i.i.d. samples, $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, estimate the sparse inverse covariance matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$.

Two approaches:

- Graphical Lasso
- CLIME

Graphical lasso

Key idea: regularizing the MLE by imposing ℓ_1 regularization (Yuan & Lin'07; Friedman, Hastie, &Tibshirani '08).

Graphical Lasso (GLasso)

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

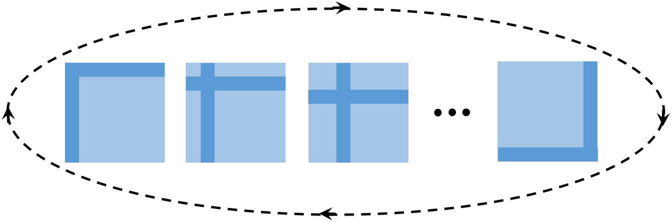
- It is a convex program! (homework)
- First-order optimality condition

$$\mathbf{\Theta}^{-1} - \boldsymbol{S} - \lambda \underbrace{\partial\|\mathbf{\Theta}\|_1}_{\text{subgradient}} = \mathbf{0} \tag{9.1}$$

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

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 working version of $\boldsymbol{\Theta}^{-1}$. Partition all matrices into 1 column/row vs. the rest

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

Blockwise coordinate descent

Blockwise step: suppose we fix all but the last row / column. It follows from (9.1) that

$$\mathbf{0} \in \boldsymbol{W}_{11}\boldsymbol{\beta} - \boldsymbol{s}_{12} + \lambda \partial \|\boldsymbol{\theta}_{12}\|_1 = \boldsymbol{W}_{11}\boldsymbol{\beta} - \boldsymbol{s}_{12} + \lambda \partial \|\boldsymbol{\beta}_{12}\|_1 \qquad (9.2)$$

where $\boldsymbol{\beta} = -\boldsymbol{\theta}_{12} \cdot w_{22}$ (by matrix inverse formula)

This coincides with optimality condition for

$$\text{minimize}_{\boldsymbol{\beta}} \quad \frac{1}{2}\|\boldsymbol{W}_{11}^{1/2}\boldsymbol{\beta} - \boldsymbol{W}_{11}^{-1/2}\boldsymbol{s}_{12}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \qquad (9.3)$$

Blockwise coordinate descent

Algorithm 9.1 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

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

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

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

Blockwise coordinate descent

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

Lemma 9.3 (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

Proof of Lemma 9.3

A key observation for the proof of Lemma 9.3

Fact 9.4 (Lemma 2, Mazumder & Hastie, '12)

Solving (9.3) is equivalent to solving

$$\text{minimize}_{\boldsymbol{\gamma}} \ (\boldsymbol{s}_{12} + \boldsymbol{\gamma})^\top \boldsymbol{W}_{11}^{-1} (\boldsymbol{s}_{12} + \boldsymbol{\gamma}) \quad \text{s.t. } \|\boldsymbol{\gamma}\|_\infty \leq \lambda \tag{9.4}$$

where solutions to 2 problems are related by $\hat{\boldsymbol{\beta}} = \boldsymbol{W}_{11}^{-1}(\boldsymbol{s}_{12} + \hat{\boldsymbol{\gamma}})$

- Check that optimality condition of (9.3) and that of (9.4) match

Proof of Lemma 9.3

Suppose in t^{th} iteration one has $\|\boldsymbol{W}^{(t)} - \boldsymbol{S}\|_\infty \le \lambda$ and

$$\boldsymbol{W}^{(t)} \succ \mathbf{0}$$

$$\iff \quad \boldsymbol{W}_{11}^{(t)} \succ \mathbf{0}; \quad w_{22} - \boldsymbol{w}_{12}^{(t)\top}\left(\boldsymbol{W}_{11}^{(t)}\right)^{-1}\boldsymbol{w}_{12}^{(t)} > 0 \quad \text{(Schur complement)}$$

We only update $\boldsymbol{w}_{12}$, so it suffices to show

$$w_{22} - \boldsymbol{w}_{12}^{(t+1)\top}\left(\boldsymbol{W}_{11}^{(t)}\right)^{-1}\boldsymbol{w}_{12}^{(t+1)} > 0 \tag{9.5}$$

Recall that $\boldsymbol{w}_{12}^{(t+1)} = \boldsymbol{W}_{11}^{t}\boldsymbol{\beta}^{t+1}$. It follows from Fact 9.4 that and

$$\|\boldsymbol{w}_{12}^{(t+1)} - \boldsymbol{s}_{12}\|_\infty \le \lambda;$$
$$\boldsymbol{w}_{12}^{(t+1)\top}(\boldsymbol{W}_{11}^{(t)})^{-1}\boldsymbol{w}_{12}^{(t+1)} \le \boldsymbol{w}_{12}^{(t)\top}(\boldsymbol{W}_{11}^{(t)})^{-1}\boldsymbol{w}_{12}^{(t)}.$$

Since $w_{22} = s_{22} + \lambda$ remains unchanged, we establish (9.5).

CLIME

Key idea: Utilize two facts:

- $\boldsymbol{\Sigma} \cdot \boldsymbol{\Theta} = \boldsymbol{I}$.
- The SCM $\boldsymbol{S}$ can be used as a surrogate of $\boldsymbol{\Sigma}$.

CLIME (Cai, Liu & Luo, 2011)

$$\text{minimize}_{\boldsymbol{\Theta}} \ \|\boldsymbol{\Theta}\|_1 \quad \text{s.t.} \quad \|\boldsymbol{S}\boldsymbol{\Theta} - \boldsymbol{I}\|_\infty \leq \lambda_n.$$

- Note: $\|\boldsymbol{A}\|_\infty = \max_{i,j} |A_{i,j}|$.
- Parallelizable for each column of $\boldsymbol{\Theta}$, thus very efficient.
- Post-processing step needed to guarantee symmetry and PSD.

Comparison with GLasso

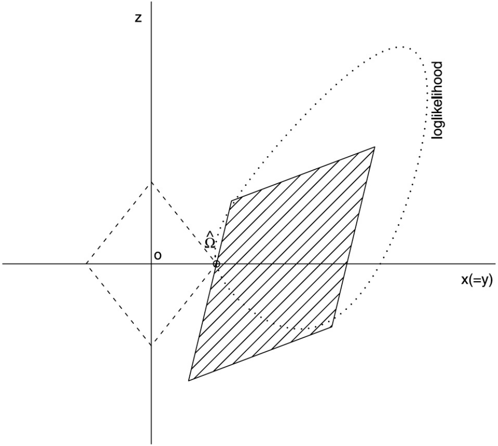


Figure 1. Plot of the elementwise ℓ_∞ constrained feasible set (shaded polygon) and the elementwise ℓ_1 norm objective (dashed diamond near the origin) from CLIME. The log-likelihood function as in Glasso is represented by the dotted line.

Figure credit: Cai, Liu & Luo, 2011.

Gaussian Graphical Model with Latent Variables

Latent variables in graphical models

Motivation: some of the variables are not directly observable.

medical/biological

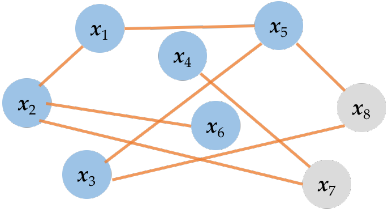
economy

We call the unobserved/missing variables the **latent variables**.

Graphical models with latent variables

What if one only observes a subset of variables?

$$\begin{bmatrix} \boldsymbol{x}_{\mathrm{o}} \\ \boldsymbol{x}_{\mathrm{h}} \end{bmatrix} \quad \begin{matrix} \text{(observed variables)} \\ \text{(hidden variables)} \end{matrix}$$



$$\boldsymbol{x}_{\mathrm{o}} = [x_1, \cdots, x_6]^\top, \boldsymbol{x}_{\mathrm{h}} = [x_7, x_8]^\top$$

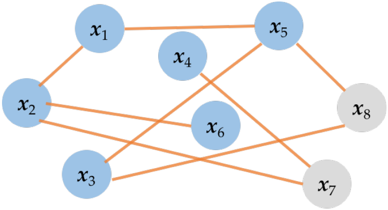
Covariance and precision matrices can be partitioned as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \overbrace{\boldsymbol{\Sigma}_{\mathrm{o}}}^{\text{observed part}} & \boldsymbol{\Sigma}_{\mathrm{o,h}} \\ \boldsymbol{\Sigma}_{\mathrm{o,h}}^\top & \boldsymbol{\Sigma}_{\mathrm{h}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Theta}_{\mathrm{o}} & \boldsymbol{\Theta}_{\mathrm{o,h}} \\ \boldsymbol{\Theta}_{\mathrm{o,h}}^\top & \boldsymbol{\Theta}_{\mathrm{h}} \end{bmatrix}^{-1}$$

Graphical models with latent variables

What if one only observes a subset of variables?

$$\begin{bmatrix} \boldsymbol{x}_{\rm o} \\ \boldsymbol{x}_{\rm h} \end{bmatrix} \quad \begin{matrix} \text{(observed variables)} \\ \text{(hidden variables)} \end{matrix}$$



$$\boldsymbol{x}_{\rm o} = [x_1, \cdots, x_6]^\top, \boldsymbol{x}_{\rm h} = [x_7, x_8]^\top$$

$$\boldsymbol{\Theta}_{\rm o} = \underbrace{\boldsymbol{\Sigma}_{\rm o}^{-1}}_{\text{observed}} = \underbrace{\boldsymbol{\Theta}_{\rm o}}_{\text{sparse}} - \underbrace{\boldsymbol{\Theta}_{\rm o,h}\boldsymbol{\Theta}_{\rm h}^{-1}\boldsymbol{\Theta}_{\rm h,o}}_{\text{low-rank if \# latent vars is small}}$$

sparse + low-rank decomposition

Inverse covariance estimation for LVGGM

Problem definition: Given n i.i.d. samples, $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, estimate the sparse - low-rank inverse covariance matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$.

First write

$$\boldsymbol{\Theta} = \boldsymbol{\Psi} - \boldsymbol{L}$$

where $\boldsymbol{\Psi} \succeq \boldsymbol{0}$, $\boldsymbol{L} \succeq \boldsymbol{0}$.

LVGGM (Chandrasekaran, Parrilo, Willsky, 2012)

$$\begin{aligned} \text{maximize}_{\boldsymbol{\Phi}, \boldsymbol{L}} \quad & \underbrace{\log\det(\boldsymbol{\Theta}) - \mathrm{tr}(\boldsymbol{S}(\boldsymbol{\Phi} - \boldsymbol{L}))}_{\text{log-likelihood}} - \lambda_n \left(\|\boldsymbol{\Psi}\|_1 + \eta \mathrm{tr}(\boldsymbol{L}) \right) \\ \text{s.t.} \quad & \boldsymbol{\Phi} - \boldsymbol{L} \succeq \boldsymbol{0}, \quad \boldsymbol{L} \succeq \boldsymbol{0}. \end{aligned}$$

Reference

[1] "*Sparse inverse covariance estimation with the graphical lasso*," J. Friedman, T. Hastie, and R. Tibshirani, *Biostatistics*, 2008.

[2] "*The graphical lasso: new insights and alternatives*," R. Mazumder and T. Hastie, *Electronic journal of statistics*, 2012.

[3] "*Statistical learning with sparsity: the Lasso and generalizations*," T. Hastie, R. Tibshirani, and M. Wainwright, 2015.

[4] "*A constrained ℓ_1 minimization approach to sparse precision matrix estimation*," T. T. Cai, W. Liu, and X. Luo, *JASA*, 2011.

[5] "*Latent variable graphical model selection via convex optimization*," V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, *The Annals of Statistics*, 2012.