Sparse Covariance Estimation with the Graphical Lasso

SHO INABA, COMPUTATIONAL SCIENCE PROGRAM (PHD).

UNIVERSITY OF MASSACHUSETTS BOSTON

Outline

1. **Introduction**
   Graphical Models
2. **Gaussian Graphical Models**
   Maximum Likelihood Estimator
3. **Graphical LASSO**
   Block Coordinate Descent for Lasso
4. **Simulation**
   Generate Graphical Models

Graphical Models

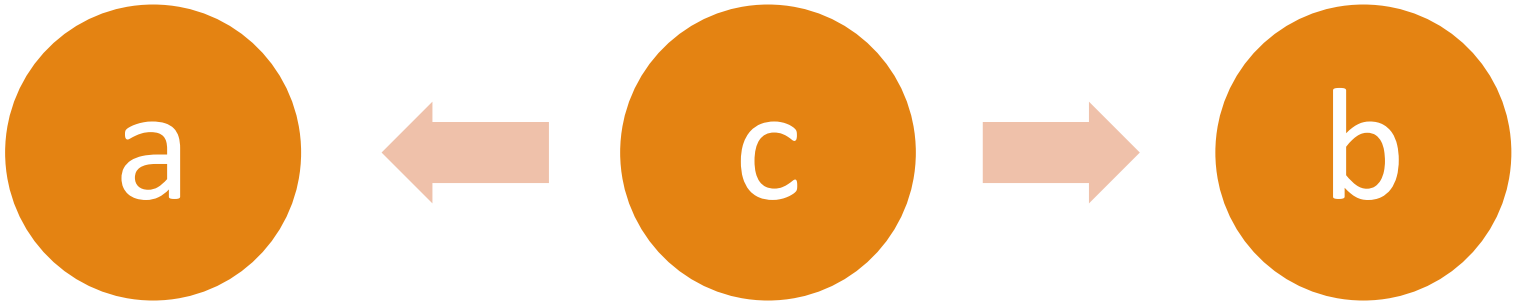
BAYES PROBABILITY AND INDEPENDENCY

What is Graphical Models?

- **Graphical representation of dependency between random variables**

Ex) Given three random variables, a, b and c, having a relationship

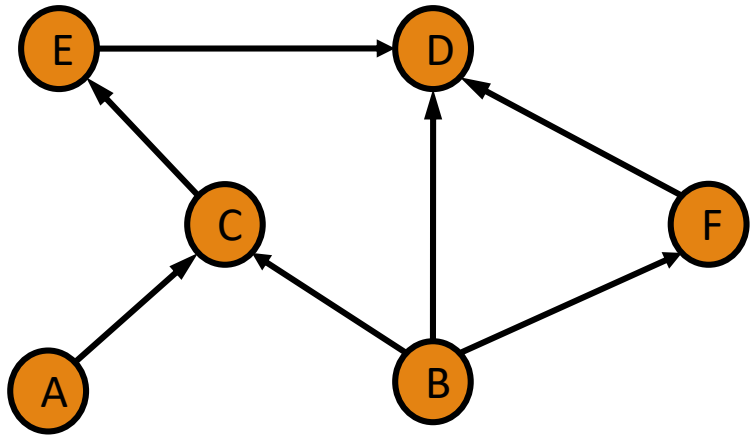
$$P(a, b, c) = P(a|c)P(b|c)P(c)$$



Two Major Graphical Models

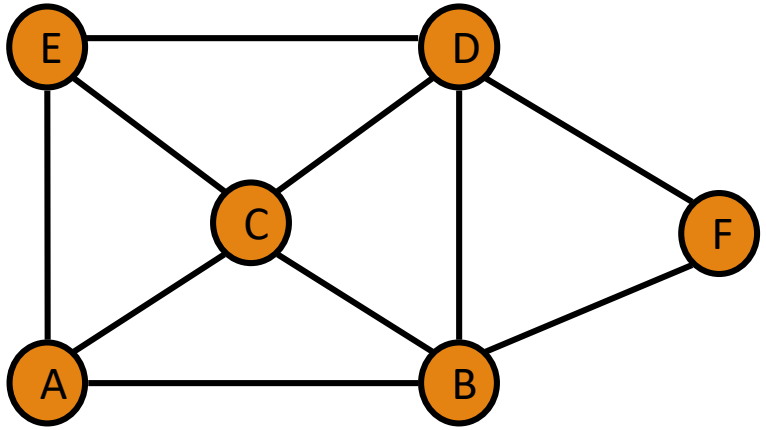
BAYESIAN NETWORK

Directed Graphical Models (DGM)



MARKOV RANDOM FIELD

Undirected Graphical Models (UGM)



4 Basic Probability Rules

- **Sum Rule**

$$p(x) = \sum_{y} p(x, y)$$

- **Product Rule**

$$p(x, y) = p(x|y)P(y)$$

- **Bayes' Theorem**

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x} P(x, y)}$$

- **Independence Rule**

$x \perp\!\!\!\perp y$ if and only if $P(x, y) = P(x)P(y)$

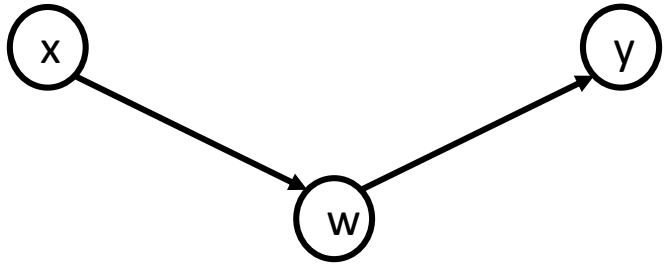
Conditional Independence

Definition $x \perp\!\!\!\perp y|w$, iff $P(x, y|w) = P(x|w)P(y|w)$

○ : Unobserved Random Variable

● : Observed Random Variable

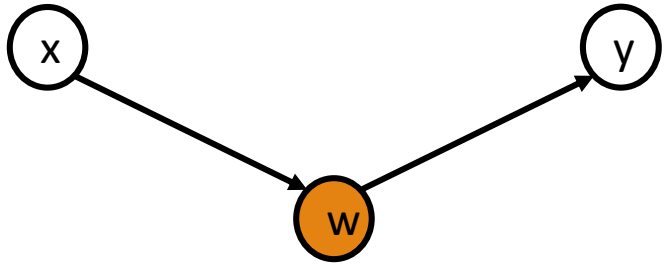
Head-to-Tail model



$P(x, w, y) = P(x)P(w|x)P(y|w)$

x, y are not independent.

Blocked model



$P(x, y|w) = P(x|w)P(y|w)$

x, y are conditionally independent.

Benefits of Graphical Models

- A complex distribution of random variables can be simplified.
- Blocked (Observed) random variable can decrease the number of necessary random variables needed to predict a set of targets.
- The model can suggest algorithms for analyzing the structure of networks.
- Visually understandable with human recognition.

Maximum Likelihood Estimator

GAUSSIAN LIKELIHOOD AND POSITIVE DEFINITE COMPLETION

Gaussian Graphical Models (GGM)

A Graphical Model with a constraint that random variables are **normally distributed**.

$$f_{\mu,\Sigma}(x) = (2\pi)^{-\frac{p}{2}} (\det\Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, x \in \mathbb{R}^p$$

With $x_i \perp\!\!\!\perp x_j$ if and only if $\Sigma_{ij} = 0$.

- For the generated Graph $G = (V, E)$, v_i and v_j are **disconnected** if $\Sigma_{ij} = 0$.

Conditional Independence of GGM

Let $S \subseteq [p] \backslash \{i, j\}$ be indices of observed random variables.

Then, the following statements are equivalent:

(a) $x_i \perp\!\!\!\perp x_j \mid x_S$;

(b) $\det\left(\Sigma_{iS,jS}\right) = 0$, where $iS = \{i\} \cup S$;

(c) $\det\left(\Sigma^{-1}_{iR,jR}\right) = 0$, where $R = [p] \backslash (S \cup \{i, j\})$.

- There is **no edge** connecting v_i and v_j if $\Sigma^{-1}_{ij} = 0$.

Maximum Likelihood Estimator Θ

Remind: $f_{\mu,\Sigma}(x) = (2\pi)^{-\frac{p}{2}} (\det\Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$

This follows the log-likelihood function $l(\mu, \Sigma|x)$ is:

$$l(\mu, \Sigma|x) = -\frac{p}{2}\ln 2\pi + \frac{1}{2}\ln\det\Sigma^{-1} - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

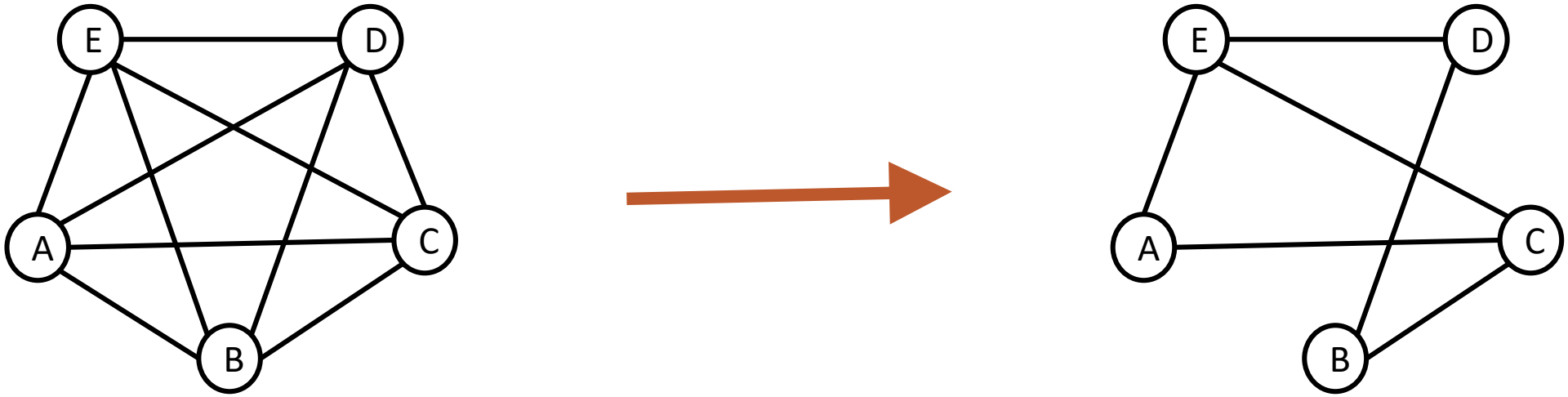
$$l(\mu, \Sigma|D) \propto \ln\det\Sigma^{-1} - \mathrm{tr}(S\Sigma^{-1}) - (\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)$$

- The optimal value for $\Sigma^{-1} = \Theta$ is driven by

$$\arg\max_{\Theta\succ 0} \ln\det\Theta - \mathrm{tr}(S\Theta)$$

Generated Graphical Model of Θ

The generated model tends to be **dense** because of the random noises.



- We need a **regularization** for obtaining a **sparse** graphical model.

Block Coordinate Descent for Lasso

REPEATED LINEAR OPTIMIZATIONS

Graphical Lasso

Add the Lasso (L_1) regularization term to the original problem:

$$\arg\max_{\Theta \succ 0} \mathrm{J}(\Theta) = \arg\max_{\Theta \succ 0} \ln\det\Theta - \mathrm{tr}(S\Theta) - \rho\|\Theta\|_1$$

Where $\|\Theta\|_1 = \sum_{ij}|\theta_{ij}|$.

- L_1 regularization provides **sparsity** into its Graphical Model.

Notations (Block Separations of Matrix)

$$\Sigma = \begin{pmatrix} \tilde{\Sigma} & \tilde{\sigma} \\ \tilde{\sigma}^T & \sigma_p \end{pmatrix}, S = \begin{pmatrix} \tilde{S} & \tilde{s} \\ \tilde{s}^T & s_p \end{pmatrix} \text{ and } \Theta = \begin{pmatrix} \tilde{\Theta} & \tilde{\theta} \\ \tilde{\theta}^T & \theta_p \end{pmatrix}$$

$\tilde{\Sigma}, \tilde{S}, \tilde{\Theta}$: $p-1 \times p-1$ submatrices of Σ, S, Θ.

$\tilde{\sigma}, \tilde{s}, \tilde{\theta}$: $p-1$ dimensional column vectors.

σ_p, s_p, θ_p: $\Sigma_{pp}, S_{pp}, \Theta_{pp}$

Algorithm

1. Σ <- $S + 2\rho$
2. n <- $\pi(p)$ (Loop starts here)
3. Permute n_ith row and column with the last row and column of Σ and S.
4. β_{n_i}<- lasso$(\tilde{\Sigma}, \tilde{s}, \rho)$, $\tilde{\sigma}$ <- $\tilde{\Sigma}\beta_{n_i}$
5. Permute back the row and column.
6. Continue until convergence. (Loop ends here)
7. θ_i <- $\frac{1}{\sigma_i - \tilde{\sigma}^T \beta_i}$, $\tilde{\theta}_i$ <- $-\theta_i \beta_i$

Gradient Descent for each element

$$\frac{\partial J(\Theta)}{\partial \Theta_{ij}} = \frac{1}{2}[\Sigma - S]_{ij} \pm \rho$$

This implies the algorithm of graphical lasso solve for an optimal covariance matrix Σ, not a precision matrix Θ.

- Since the diagonal elements of positive definite matrix is positive,

$$\Sigma_{ii} = S_{ii} + 2\rho$$

Block Coordinate Descent

Remind: we want to find an optimal $\Sigma = \begin{pmatrix} \tilde{\Sigma} & \tilde{\sigma} \\ \tilde{\sigma}^T & \sigma_p \end{pmatrix}$.

Fix the diagonal blocks $\tilde{\Sigma}$ and σ_p, then optimize $\tilde{\sigma}$.

Then, the optimization problem is equivalent to the duality:

$$\arg\min_{\beta} \tilde{J}(\beta) = \arg\min_{\beta} \frac{1}{4}\beta^T\tilde{\Sigma}\beta - \frac{1}{2}\tilde{s}^T\beta + \rho\|\beta\|_1$$

Where $\beta = \tilde{\Sigma}^{-1}\tilde{\sigma}$.

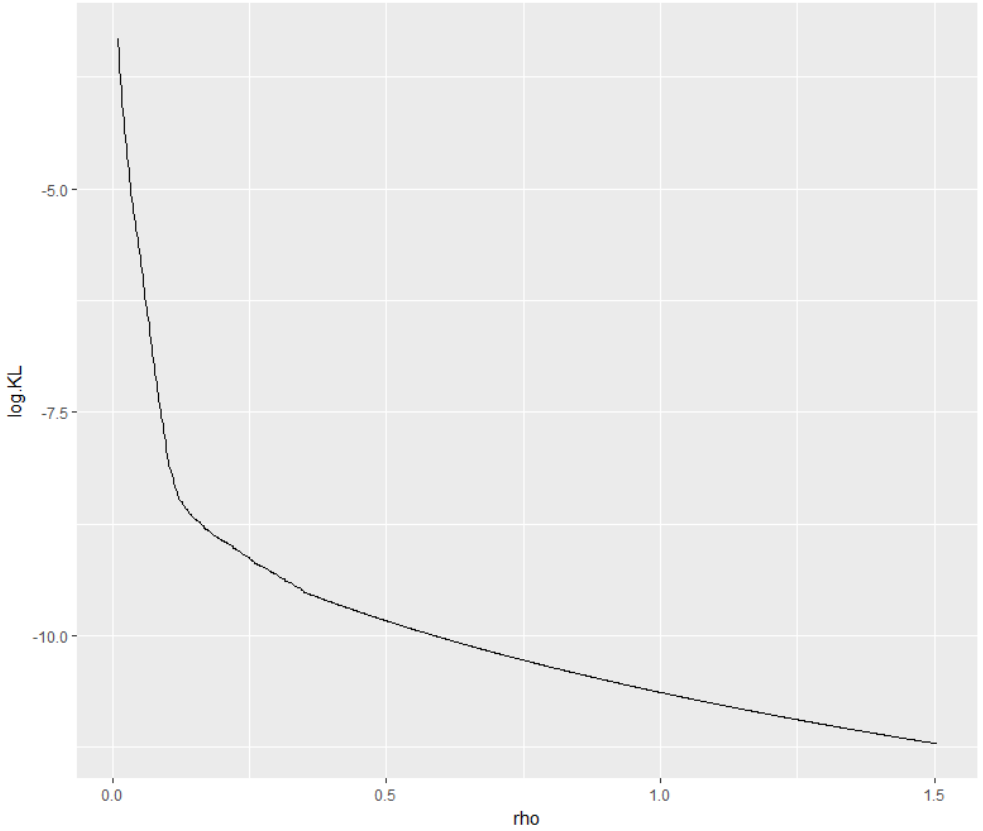
Simulations

HOW TO USE GRAPHICAL LASSO IN R

Methods

- 100 random variables
- 20 edges are randomly chosen to be connected
- Create random correlations between those dependent variables
- Compute its inverse and generate 200 random samples
- Compute a sample covariance
- Find optimal values of ρ with using KL distance
- Estimate a precision matrix and compare the intersections

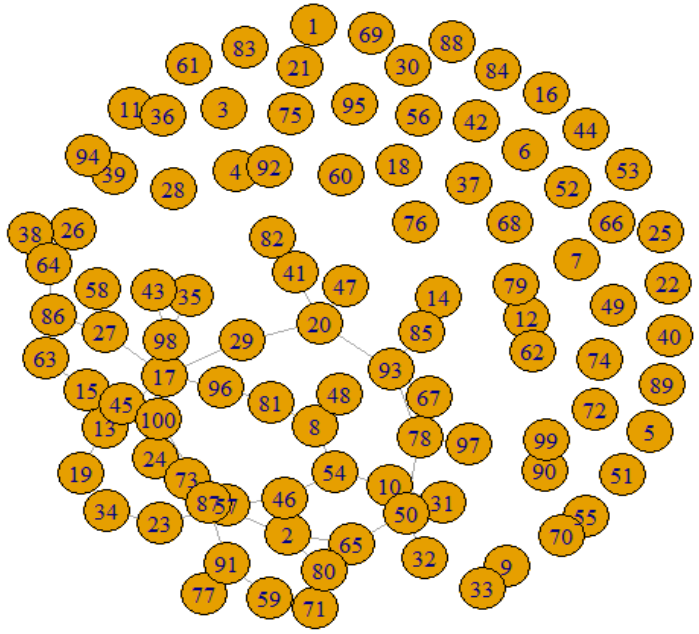
Result (KL distance of two successive Θ)



Optimal ρ is about 0.1.

The generated graph would be partially connected when ρ is between 0.1 and 0.3.

Result (Generated Graphical Model)



Roughly 40 are connected.

The intersection of connections between true value and this model is 16/20.

Reference

- Caroline Uhler; Gaussian Graphical Models: An Algebraic and Geometric Perspective, *ArXiv e-prints*, July 2017, http://adsabs.harvard.edu/abs/2017arXiv170704345U

- Jerome Friedman, Trevor Hastie, Robert Tibshirani; Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, Volume 9, Issue 3, 1 July 2008, Pages 432–441, https://doi.org/10.1093/biostatistics/kxm045