Chapter 17: Undirected Graphical Models

The Elements of Statistical Learning

Biaobin Jiang

Department of Biological Sciences
Purdue University

bjiang@purdue.edu

October 30, 2014

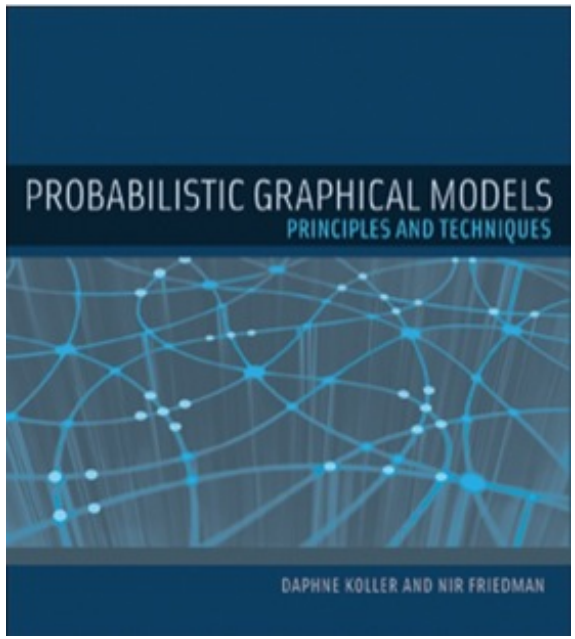
Overview

1. Introduction
   - Probabilistic Graphical Models
   - Review: Multivariate Statistics
   - Review: Matrix Operations
2. Undirected Graphical Models for Continuous Variables
   - Connection with Multiple Linear Regression
   - Estimation of Parameters with Known Structure
   - Estimation of Graph Structure
3. Undirected Graphical Models for Discrete Variables

What is Probabilistic Graphical Models

A graph consists of a set of vertices (nodes), along with a set of edges joining some pairs of the vertices.

In graphical models, each vertex represents a random variable, and the graph gives a visual way of understanding the joint distribution of the entire set of random variables.



How it works

Categories of PGM

- Directed Graphical Models, a.k.a. Bayesian Networks
- Undirected Graphical Models, a.k.a. Markov Random Field

Computational Tasks of PGM

- *Structuring*, choosing the structure of the graph;
- *Learning*, estimating the edge parameters from data; and
- *Inference*, computing marginal vertex probabilities and expectations from their joint distribution.

MultiVariate Normal Distribution

The MVN distribution is a generalization of the univariate normal distribution which has the density function (p.d.f.)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

where μ is mean of distribution, σ^2 is variance. In p-dimensions the density becomes

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right\}$$

where $\boldsymbol{\mu}$ is a p-dimensional mean vector and Σ is a symmetric covariance matrix.

Conditional Probability of MVN

Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ be a partitioned MVN random p-vector, with mean $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

The conditional distribution of X_2 given $X_1 = x_1$ is an MVN with

$$\mathbb{E}(X_2|X_1 = x_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$$
$$\mathrm{Cov}(X_2|X_1 = x_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

Matrix Trace

In Linear Algebra, the *trace* of an n-by-n square matrix $\mathbf{A}$ is defined to be the sum of the elements on the main diagonal of $\mathbf{A}$, i.e.,

$$\mathrm{tr}(\mathbf{A}) = a_{11} + a_{22} + \cdots + a_{nn} = \sum_{i=1}^{n} a_{ii}.$$

Matrix trace has several basic properties:

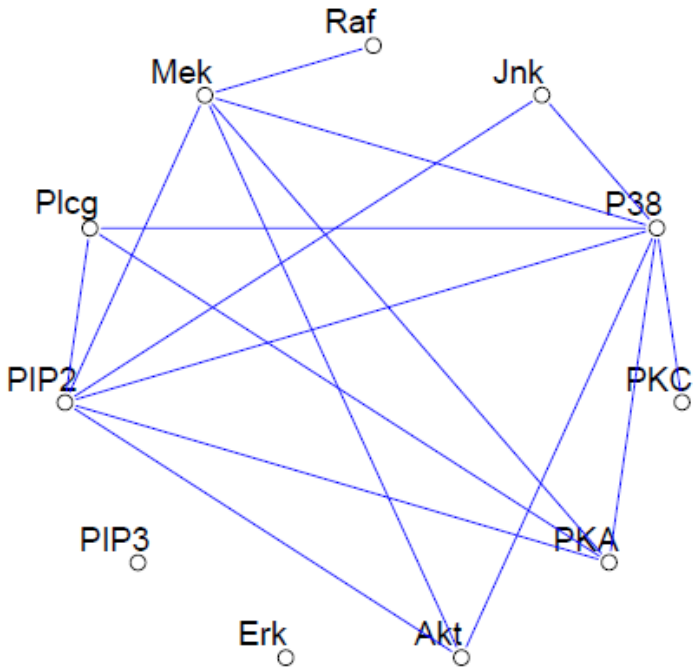
$$\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$$
$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$$

Estimation of Parameters with Known Graph Structure

What is Parameter Estimation

Given empirical covariance matrix $\mathbf{S}$, find the optimal estimation $\hat{\Sigma} = \mathbf{W}$ and its inverse $\hat{\Sigma}^{-1} = \Theta$.

In particular, if the ijth component of Θ is zero, then variable i and j are conditionally independent, given the other variables. In other words, there is no edge connection between vertex i and j.



Conditional Mean and Multiple Linear Regression

Suppose we partition $X = (Z, Y)$ where $Z = (X_1, \ldots, X_{p-1})$ and $Y = X_p$. Then we have the conditional distribution of Y given Z (Eq. (17.6))

$$(Y|Z = z) \sim \mathcal{N}(\mu_Y + (z - \mu_Z)^T \Sigma_{ZZ}^{-1} \sigma_{ZY}, \sigma_{YY} - \sigma_{ZY}^{-1} \Sigma_{ZZ}^{-1} \sigma_{ZY})$$

where we have partitioned Σ as (Eq. (17.7))

$$\Sigma = \begin{bmatrix} \Sigma_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{bmatrix}.$$

The conditional mean in Eq. (17.6) has exactly the same form as the population multiple linear regression of Y on Z, with regression coefficient $\beta = \Sigma_{ZZ}^{-1} \sigma_{ZY}$. (Proof on next page)

Proof

Given Eq. (2.9), we have expected prediction error as

$$\begin{aligned}\text{EPE}(f) &= \mathbb{E}(y - f(\mathbf{z}))^2 \\ &= \mathbb{E}(y - \mathbf{z}^T\beta)^2 \\ &= \mathbb{E}[y^2 - 2y\mathbf{z}^T\beta + \beta^T\mathbf{z}\mathbf{z}^T\beta]\end{aligned}$$

By differentiating the expected function, we have

$$\begin{aligned}\frac{d\text{EPE}(f)}{d\beta} &= \mathbb{E}\left[\frac{d(y^2 - 2y\mathbf{z}^T\beta + \beta^T\mathbf{z}\mathbf{z}^T\beta)}{d\beta}\right] \\ &= \mathbb{E}\left[-2y\mathbf{z} + 2\mathbf{z}\mathbf{z}^T\beta\right] = 0\end{aligned}$$

Then we derive $\beta = \mathbb{E}(\mathbf{z}\mathbf{z}^T)^{-1}\mathbb{E}[y\mathbf{z}] = \Sigma_{ZZ}^{-1}\sigma_{ZY}$.

How to Solve its Inverse Θ

The standard formulas for partitioned inverses give $\Sigma\Theta = \mathbf{I}$, i.e.,

$$\begin{bmatrix} \Sigma_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{bmatrix} \begin{bmatrix} \Theta_{ZZ} & \theta_{ZY} \\ \theta_{ZY}^T & \theta_{YY} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}.$$

Then we derive

$$\Sigma_{ZZ}\theta_{ZY} + \sigma_{ZY}\theta_{YY} = 0$$
$$\sigma_{ZY}^T\theta_{ZY} + \sigma_{YY}\theta_{YY} = 1$$

To solve these two equations, we have Eq. (17.8)

$$\theta_{ZY} = -\theta_{YY}\Sigma_{ZZ}^{-1}\sigma_{ZY}$$

where $1/\theta_{YY} = \sigma_{YY} - \sigma_{ZY}^T\Sigma_{ZZ}^{-1}\sigma_{ZY} > 0$. And hence, we have Eq. (17.9) $\beta = \Sigma_{ZZ}^{-1}\sigma_{ZY} = -\theta_{ZY}/\theta_{YY}$.

What We Have Learned

- The dependence of Y on Z in (17.6) is in the mean term alone. Here we see exactly that zero elements in β and hence θ_{ZY} mean that the corresponding elements of Z are conditionally independent of Y.
- We can learn about this dependence structure through Multiple Linear Regression.

Maximum Likelihood Estimation of MVN

Let $\mathbf{X}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ be sampled from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. And the MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the sample mean and empirical covariance (Eq. (17.10))

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

The likelihood function is a function of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given the data $\mathbf{X}$

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = \prod_{i=1}^{N} f(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$= (2\pi)^{-\frac{Np}{2}}|\Sigma|^{-\frac{N}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}$$

Log-likelihood

Then the log-likelihood of the data can be written as

$$\begin{aligned}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -2 \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) \\ &= N \log |\Sigma| + \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + C\end{aligned}$$

which is equivalent to Eq. (17.11) since

$$\begin{aligned}\ell(\boldsymbol{\Theta}) &= \log \det \boldsymbol{\Theta} - \text{tr}(\mathbf{S}\boldsymbol{\Theta}) \\ &= -\log |\Sigma| - \text{tr}(\sum_i (\mathbf{x}_i - \boldsymbol{\mu}) \cdot (\mathbf{x}_i - \boldsymbol{\mu})^T \Theta) \\ &= -\log |\Sigma| - \sum_i \text{tr}((\mathbf{x}_i - \boldsymbol{\mu})^T \Theta \cdot (\mathbf{x}_i - \boldsymbol{\mu})) \\ &= -\log |\Sigma| - \sum_i (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\end{aligned}$$

Missing Edges: Equality Constraints

Now, we would like to maximize the log-likelihood under the constraints that some pre-defined subset of the parameters are zero.

$$\begin{aligned} \underset{\boldsymbol{\Theta}}{\text{maximize}} \quad & \ell_C(\boldsymbol{\Theta}) = \log\det\boldsymbol{\Theta} - \text{tr}(\mathbf{S}\boldsymbol{\Theta}) \\ \text{subject to} \quad & \theta_{jk} = 0,\ (j,k) \notin \mathbf{E} \end{aligned}$$

Then we add Lagrange multiplier, and derive Eq. (17.12)

$$\underset{\boldsymbol{\Theta}}{\text{maximize}} \quad \ell_C(\boldsymbol{\Theta}) = \log|\boldsymbol{\Theta}| - \text{tr}(\mathbf{S}\boldsymbol{\Theta}) - \sum_{(j,k)\notin\mathbf{E}} \gamma_{j,k}\theta_{j,k}$$

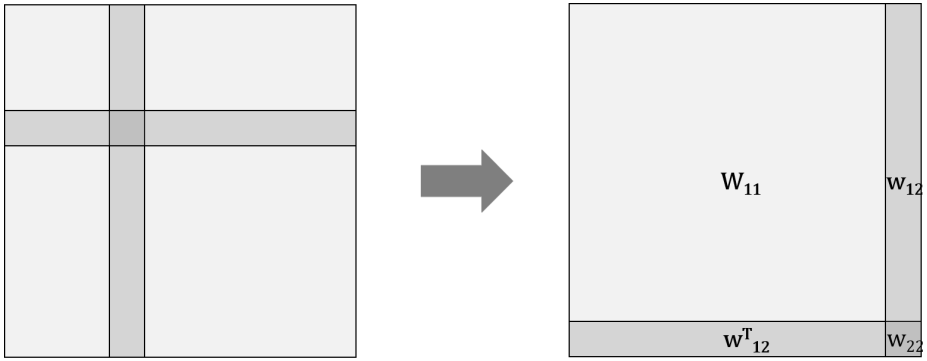
Taking the derivative, we have Eq. (17.13)

$$\boldsymbol{\Theta}^{-1} - \mathbf{S} - \boldsymbol{\Gamma} = \mathbf{0}$$

where $\boldsymbol{\Gamma}$ is a matrix of Lagrange parameters with nonzero values for all missing edges.

Solve (17.13) by Multiple Linear Regression

Step 1: Partition $\mathbf{W}$ and derive Eq. (17.14): $w_{12} - s_{12} - \gamma_{12} = 0$.



Step 2: Connect $\mathbf{w}_{12}$ with β. Eq. (17.16)

$$\begin{bmatrix} \mathbf{W}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}^T & w_{22} \end{bmatrix} \begin{bmatrix} \mathbf{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}_{12}^T & \theta_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}.$$

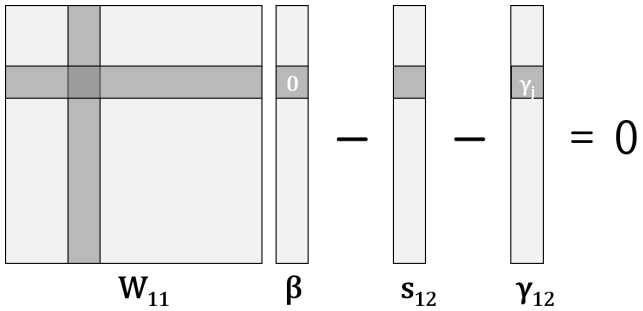
This implies Eq. (17.17)

$$\mathbf{w}_{12} = -\mathbf{W}_{11}\theta_{12}/\theta_{22} = \mathbf{W}_{11}\beta$$

Solve (17.13) by Multiple Linear Regression (Cont.)

Step 3: Use simple *subset* regression to solve Eq. (17.18)

$$\mathbf{W}_{11}\beta - \mathbf{s}_{12} - \gamma_{12} = 0$$



If $\gamma_j \neq 0$, we remove all the elements in jth row and jth column, and derive the reduced system of equation Eq. (17.19)

$$\mathbf{W}_{11}^{*}\beta^{*} - \mathbf{s}_{12}^{*} = 0$$

Step 4: Update θ_{22} and θ_{12} (Eq. (17.20))

$$1/\theta_{22} = s_{22} - \mathbf{w}_{12}^{T}\hat{\beta}$$

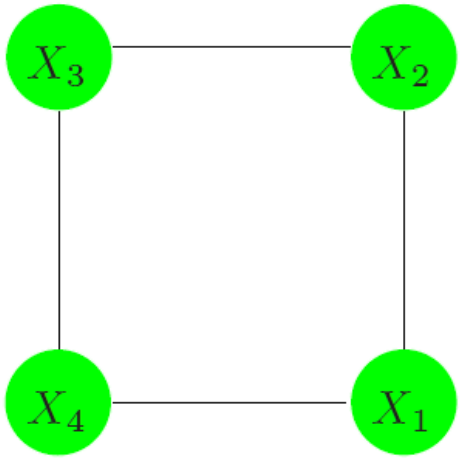
$$\theta_{12} = -\hat{\beta}\theta_{22}.$$

Summary: Algorithm 17.1

Algorithm 17.1 *A Modified Regression Algorithm for Estimation of an Undirected Gaussian Graphical Model with Known Structure.*

1. Initialize $\mathbf{W} = \mathbf{S}$.
2. Repeat for $j = 1, 2, \ldots, p$ until convergence:
   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.
   (b) Solve $\mathbf{W}_{11}^{*}\beta^{*} - s_{12}^{*} = 0$ for the unconstrained edge parameters β^{*}, using the reduced system of equations as in (17.19). Obtain $\hat{\beta}$ by padding $\hat{\beta}^{*}$ with zeros in the appropriate positions.
   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$
3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = s_{22} - w_{12}^{T}\hat{\beta}$.

A Case Study: Figure 17.4



$$\mathbf{S} = \begin{pmatrix} 10 & 1 & 5 & 4 \\ 1 & 10 & 2 & 6 \\ 5 & 2 & 10 & 3 \\ 4 & 6 & 3 & 10 \end{pmatrix}$$

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 10.00 & 1.00 & 1.31 & 4.00 \\ 1.00 & 10.00 & 2.00 & 0.87 \\ 1.31 & 2.00 & 10.00 & 3.00 \\ 4.00 & 0.87 & 3.00 & 10.00 \end{pmatrix}, \quad \hat{\mathbf{\Sigma}}^{-1} = \begin{pmatrix} 0.12 & -0.01 & 0.00 & -0.05 \\ -0.01 & 0.11 & -0.02 & 0.00 \\ 0.00 & -0.02 & 0.11 & -0.03 \\ -0.05 & 0.00 & -0.03 & 0.13 \end{pmatrix}$$

Estimation of the Graph Structure

Graph Lasso

Graph Lasso

Graph Lasso fits a lasso regression using each variable as the response and the others as predictors. Consider maximizing the penalized log-likelihood Eq. (17.21)

$$\log|\mathbf{\Theta}| - \text{tr}(\mathbf{S\Theta}) - \lambda\|\mathbf{\Theta}\|_1$$

where $\|\mathbf{\Theta}\|_1$ is the L_1 norm, i.e., the sum of the absolute values of the elements of $\mathbf{\Theta}$.

Similarly, taking the differentiation, we reach the analog of Eq. (17.18) as Eq. (17.23)

$$\mathbf{W}_{11}\beta - \mathbf{s}_{12} + \lambda \cdot \text{Sign}(\beta) = \mathbf{0}.$$

Cyclical Coordinate Descent Algorithm

Let's re-denote the following equation

$$\mathbf{W}_{11}\beta - \mathbf{s}_{12} + \lambda \cdot \text{Sign}(\beta) = \mathbf{0}.$$

as a $(p-1)$ by $(p-1)$ linear system using $\mathbf{A}, \mathbf{x}$ and $\mathbf{b}$

$$\mathbf{A}\mathbf{x} - \mathbf{b} + \lambda \cdot \text{Sign}(\mathbf{x}) = \mathbf{0}.$$

For $i = 1, 2, \ldots, p-1, 1, 2, \ldots, p-1, \ldots$, we update (Eq. (17.26))

$$x_i \leftarrow \text{St}\left(b_i - \sum_{k \neq i} \mathbf{A}_{ki} x_k, \lambda \right) / \mathbf{A}_{ii}$$

where $\text{St}(x, t)$ is the soft-threshold operator (Eq. (17.27))

$$\text{St}(x, t) = \text{sign}(x) \cdot (|x| - t)_+$$

Summary: Graph Lasso Algorithm

Algorithm 17.2 *Graphical Lasso.*

1. Initialize $\mathbf{W} = \mathbf{S} + \lambda\mathbf{I}$. The diagonal of $\mathbf{W}$ remains unchanged in what follows.

2. Repeat for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ until convergence:

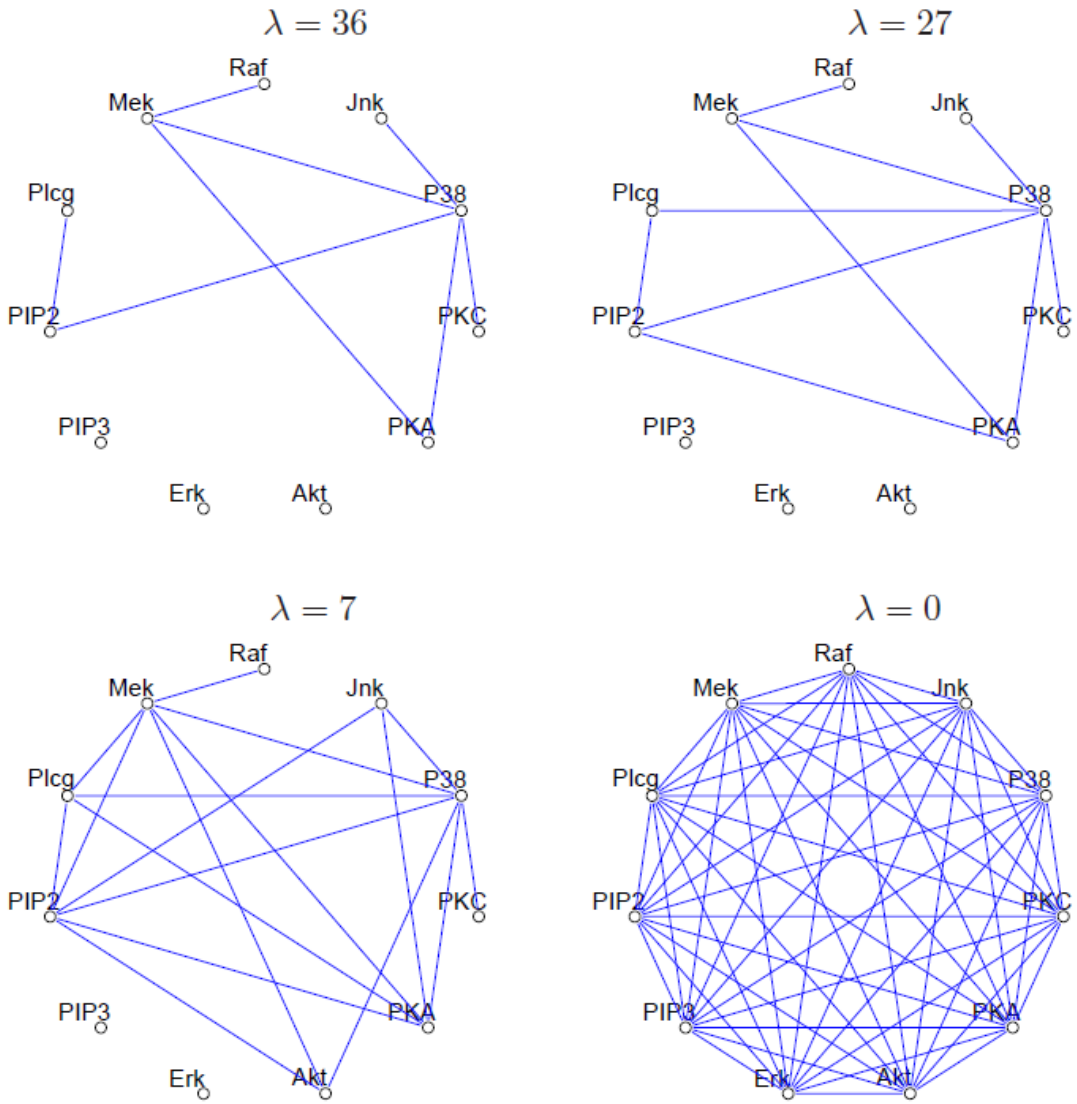
   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.

   (b) Solve the estimating equations $\mathbf{W}_{11}\beta - s_{12} + \lambda \cdot \text{Sign}(\beta) = 0$ using the cyclical coordinate-descent algorithm (17.26) for the modified lasso.

   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$

3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = w_{22} - w_{12}^T\hat{\beta}$.

A Case Study: Flow-Cytometry Data



Missing/Hidden Node Values: EM

Note that the values at some of the nodes in a graphical model can by *unobserved*; i.e., missing or hidden.

The EM algorithm can be used to impute the missing values with **E Step** (Eq. (17.43)) imputing the missing values from the current estimates of μ and Σ

$$\hat{x}_{i,m_i} = \mathbb{E}(x_{i,m_i}|x_{i,o_i},\theta) = \hat{\mu}_{m_i} + \hat{\Sigma}_{m_i,o_i}\hat{\Sigma}^{-1}_{o_i,o_i}(x_{i,o_i} - \hat{\mu}_{o_i})$$

and **M Step** (Eq. (17.44)) re-estimating μ and Σ from the empirical mean and (modified) covariance of the imputed data

$$\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N}\hat{x}_{ij}$$

$$\hat{\Sigma}_{jj'} = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_{ij} - \hat{\mu}_j)(\hat{x}_{ij'} - \hat{\mu}_{j'}) + c_{i,jj'}$$

Ising Models/Boltzmann Machines

Pairwise Markov networks with *binary* variables are called *Ising models* in statistical mechanics, and *Boltzmann machines* in machine learning.

The joint probabilities of the Ising model is given by Eq. (17.28, 17.29)

$$\mathbb{P}(X, \mathbf{\Theta}) = \frac{1}{\Phi(\mathbf{\Theta})} \prod_{C \in \mathcal{C}} \psi_C(x_C) = \frac{\exp\left(\sum_{(j,k) \in \mathbf{E}} \theta_{jk} X_j X_k \right)}{\sum_{x \in \mathcal{X}} \left[\exp\left(\sum_{(j,k) \in \mathbf{E}} \theta_{jk} x_j x_k \right) \right]}$$

The Ising model implies a logistic form for each node conditional on the other (Eq. (17.30))

$$\mathbb{P}(X_j = 1 | X_{-j} = x_{-j}) = \frac{1}{1 + \exp\left(-\theta_{j0} - \sum_{(j,k) \in \mathbf{E}} \theta_{jk} x_k \right)}$$

where X_{-j} denotes all of the nodes except j.

Estimation of Parameters with Known Graph Structure

Given X, find $\mathbf{\Theta}$.

The log-likelihood is Eq. (17.31)

$$\ell(\mathbf{\Theta}) = \sum_{i=1}^{N} \log \mathbb{P}_{\mathbf{\Theta}}(X_i = x_i) = \sum_{i=1}^{N} \left[\sum_{(j,k)\in \mathbf{E}} \theta_{jk} x_{ij} x_{ik} - \Phi(\mathbf{\Theta}) \right]$$

The gradient of the log-likelihood is Eq. (17.32, 17.33, 17.34)

$$\begin{aligned} \frac{\partial \ell(\mathbf{\Theta})}{\partial \theta_{jk}} &= \sum_{i=1}^{N} x_{ij} x_{ik} - N \sum_{x\in\mathcal{X}} x_j x_k \cdot p(x, \mathbf{\Theta}) \\ &= \hat{\mathbb{E}}(X_j X_k) - \mathbb{E}_{\mathbf{\Theta}}(X_j X_k) \\ &= 0 \end{aligned}$$

Reference

- T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning*.
- D. Koller, N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*.

The End