Notes on Statistical and Machine Learning

Graphical Models

Chapter: *30* **Prepared by:** *Chenxi Zhou*

This note is prepared based on

- *Chapter 8, Graphical Models* in Bishop (2016), and
- *Chapter 17, Undirected Graphical Models* in Hastie, Tibshirani, and Friedman (2009).

I. Introduction

1. **Why Graphical Models?:** *Probabilistic graphical models*, or simply *graphical models*, use diagrams to represent probability distributions.

   They have the following advantages:

   (a) They provide a simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models;

   (b) Insights into the properties of the model, including conditional independence properties, can be obtained by inspection of the graph;

   (c) Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along implicitly.

2. **Graph:** A *graph*, denoted by $\mathcal{G}$, consists of a pair (V, E), where V is a set of vertices (node) and E is a set of edges. In a probabilistic graphical model,

   (a) each vertex represents a random variable (or group of random variables), and

   (b) the edges express probabilistic relationships between these variables.

3. **Directed Graphical Models:** In a *directed graphical model*, also known as *Bayesian networks*, the edges in the graph have a particular directionality indicated by arrows.

   Remark. Directed graphical models represent probability distributions that can be factored into products of conditional distributions, and have the potential for *causal interpretations*.

4. **Undirected Graphical Models:** In an *undirected graph*, also known as *Markov random fields*, the edges have *no* directional arrows.

   Remark. The absence of an edge between two vertices means that the corresponding random variables are conditionally independent, given the other variables.

5. **Challenges in Graphical Models:** The *main challenges* in working with graphical models are

   (a) model selection (choosing the structure of the graph),

   (b) estimation of the edge parameters from data, and

   (c) computation of marginal vertex probabilities and expectations from their joint distribution.

   Remark. The last two tasks are called *learning* and *inference*, respectively, in the computer science literature.

II. Directed Graphical Model

II.1 Introduction

1. **Motivating Example:** Consider the random variables X, Y and Z whose joint probability density function (pdf) is given by $f_{X,Y,Z}$. Using the product rule of probability, we have

$$f_{X,Y,Z}(x,y,z) = f_{Z|X,Y}(z \,|\, x,y) f_{Y|X}(y \,|\, x) f_X(x), \tag{1}$$

   where $f_{Z|X,Y}$ is the conditional pdf of Z given $X = x$ and $Y = y$, $f_{Y|X}$ is the conditional pdf of Y given $X = x$, and f_X is the marginal pdf of X.

   In order to create a diagram associated with the factorization (1), we do the following:

   (a) Draw a vertex associated with each random variable;

   (b) For each conditional distribution, we add a directed edge (arrow) to the graph from the vertex corresponding to the variables on which the distribution is conditioned.

   The resulting diagram is shown in Figure 1.

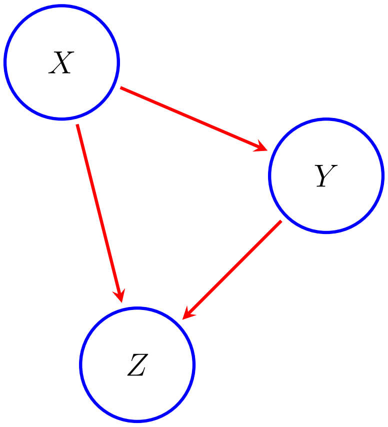

Figure 1: Diagram associated with the factorization (1).

2. **Parent and Child Vertices:** If there is a directed edge going from a vertex A to a vertex B, then we say that

   (a) the vertex A is the *parent* of the vertex B, and

   (b) the vertex B is the *child* of vertex A.

3. **Fully Connected Graph:** A graph is said to be *full connected* if there is a link between every pair of vertices.

4. **Relationship between a Given Directed Graph and Distribution of Random Variables:**

   (a) The joint distribution defined by a graph is given by the product, over all of the nodes of the graph, of a conditional distribution for each node conditioned on the variables corresponding to the *parents* of that node in the graph. Mathematically, for a graph with m nodes, the joint distribution can be factored as

   $$f_{X_1, X_2, \cdots, X_m}(x_1, x_2, \cdots, x_m) = \prod_{j=1}^{m} f_{X_j \mid \mathrm{pa}_j}(x_j \mid \mathrm{pa}_j), \tag{2}$$

   where pa_j denotes the set of *parents* of x_j.

   Remark. Equation (2) expresses the *factorization properties* of the joint distribution for a directed graphical model.

   (b) Conversely, given the factorization in the form of (2), we can associate sets of variables with the nodes of a directed graph.

5. **Directed Acyclic Graphs:** A graph is said to be a *directed acyclic graph* if it contains no directed cycles, i.e., there are *no* closed paths within the graph such that we can move from node to node along links following the direction of the arrows and end up back at the starting node.

   Remark. A directed acyclic graph is equivalent to that there exists an ordering of the nodes such that there are *no* links that go from any node to any lower numbered node.

6. **Practical Applications of Directed Acyclic Graphs:** Typically,

   (a) the higher-numbered variables correspond to terminal nodes of the graph that represent the *observations*,

   (b) the lower-numbered variables correspond to latent variables.

   The primary role of the latent variables is to allow a complicated distribution over the observed variables to be represented in terms of a model constructed from simpler conditional distributions.

7. **Ancestral Sampling:**

   (a) *Setup:* Consider a joint pdf over m random variables factorized according to (2) corresponding to a directed acyclic graph. Suppose the m variables have been ordered such that there are no links from any node to any lower numbered node.

(b) *Goal:* Our goal is to draw a sample $X_1, X_2, \cdots, X_m$ from the joint distribution.

(c) *Procedure:*

i. Start from the lowest-numbered node and draw a sample from the distribution f_{X_1}, which we call X_1;

ii. Work through each of the nodes in order, so that for node j, we draw a sample from the conditional distribution $f_{X_j|\mathrm{pa}_j}$ in which the parent variables have been set to their sampled values.

iii. Continue the preceding step until $j = m$.

(d) *Remarks:*

i. Note that, at each stage, the parent values will *always* be available because they correspond to lower-numbered nodes that have already been sampled.

ii. To obtain a sample from some *marginal* distribution corresponding to a subset of the random variables, we simply take the sampled values for the required nodes and *ignore* the sampled values for the remaining nodes.

II.2 Discrete Random Variables

1. **Setup:** We assume all nodes correspond to discrete random variables.

2. **Univariate Discrete Random Variable:** We consider a single random variable that can take on K distinct values. The corresponding pmf can be written as

$$\mathbb{P}(X = \mathbf{x}) = \prod_{k=1}^{K} \mu_k^{x_k},$$

which is governed by the parameter $\boldsymbol{\mu} := (\mu_1, \mu_2, \cdots, \mu_K)^\top \in \mathbb{R}^K$ with $\sum_{k=1}^{K} \mu_k = 1$. In addition, $\mathbf{x} := (x_1, x_2, \cdots, x_K)^\top \in \mathbb{R}^K$ and exactly one component is equal to 1 and all others are 0.

Number of parameters to be determined: Due to the unity constraint, only $K-1$ values for μ_k need to be specified in order to define the distribution.

3. **Bivariate Discrete Random Variable:** Suppose we have two discrete random variables X_1 and X_2, each of which can take on K different values.

Let the probability of observing exactly the k-th component in X_1 equal to 1 and exactly the ℓ-th component in X_2 equal to 1 be $\mu_{k,\ell}$. The joint pmf of X_1 and X_2 is

$$\mathbb{P}(X_1 = \mathbf{x}_1, X_2 = \mathbf{x}_2) = \prod_{k=1}^{K} \prod_{\ell=1}^{K} \mu_{k,\ell}^{x_{1,k} x_{2,\ell}},$$

where $\mathbf{x}_1 = (x_{1,1}, x_{1,2}, \cdots, x_{1,K})^\top \in \mathbb{R}^K$ and $\mathbf{x}_2 = (x_{2,1}, x_{2,2}, \cdots, x_{2,K})^\top \in \mathbb{R}^K$.

Number of parameters to be determined: Since the parameters $\{\mu_{k,\ell}\}_{k,\ell}$ must satisfy $\sum_{k=1}^{K} \sum_{\ell=1}^{K} \mu_{k,\ell} = 1$, the distribution is governed by $K^2 - 1$ parameters in total.

4. **Bivariate Discrete Random Variable — Case 1:** We can rewrite the joint pmf of X_1 and X_2 as

$$\mathbb{P}(X_1 = \mathbf{x}_1, X_2 = \mathbf{x}_2) = \mathbb{P}(X_2 = \mathbf{x}_2 \,|\, X_1 = \mathbf{x}_1)\mathbb{P}(X_1 = \mathbf{x}_1),$$

corresponding to a two-node graph with a link going from the X_1 node to the X_2 node (see Figure 2).

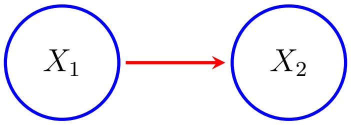

Figure 2: Diagram associated with bivariate random variable case where X_2 is conditional on X_1.

The marginal distribution f_{X_1} is governed by $K-1$ parameters. The conditional distribution $f_{X_2|X_1}$ requires the specification of $K-1$ parameters for *each* of the K possible values of X_1. The total number of parameters that must be specified in the joint distribution is then

$$(K-1) + K \times (K-1) = K^2 - 1.$$

5. **Bivariate Discrete Random Variable — Case 2:** We assume the random variables X_1 and X_2 are independent, corresponding to the graphical model shown in (3).

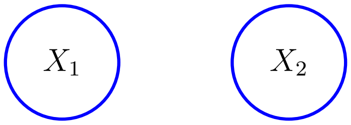

Figure 3: Diagram associated with bivariate random variable case where X_1 and X_2 are independent.

Then, the total number of parameters is $2 \times (K-1)$.

6. **m-variate Discrete Random Variable:** We further extent to the case of m discrete random variables, and assume each of m random variables can take on K distinct values.

   (a) If we do *not* put any constraints on the distribution and the corresponding directed acyclic graph is fully connected, the total number of parameters we need to determine to specify the distribution is $K^m - 1$.

   (b) If we assume that $X_1, X_2, \cdots, X_m$ are all independent, the total number of parameters we need to determine to specify the distribution is $m(K-1)$.

   (c) We can also make some assumptions on the independence among $X_1, X_2, \cdots, X_m$ so that the resulting graph is between full connectivity and full independence. As

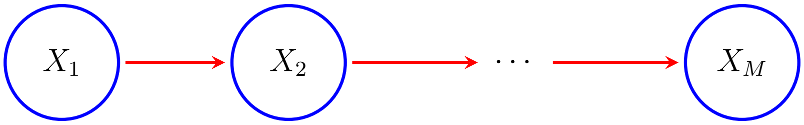

Figure 4: A special case of M-variate discrete random variable.

one example, assume the joint pmf of $X_1, X_2, \cdots, X_m$ can be factorized as

$$\begin{aligned}\mathbb{P}(X_1 = \mathbf{x}_1, X_2 = \mathbf{x}_2, \cdots, X_m = \mathbf{x}_m) \\ = \mathbb{P}(X_1 = \mathbf{x}_1) \prod_{j=2}^{m} \mathbb{P}(X_j = \mathbf{x}_j \,|\, X_{j-1} = \mathbf{x}_{j-1}).\end{aligned}$$

The corresponding graph is given by Figure 4. In this case,

i. the total number of parameters to determine the distribution of X_1 is $K-1$,
ii. the total number of parameters to determine the distribution of X_j given X_{j-1} is $K(K-1)$,

so that the total number of parameters to determine the distribution of $X_1, X_2, \cdots, X_m$ is

$$(K-1) + (m-1)K(K-1) = mK^2 +,$$

which grows *quadratically* in K and grows *linearly* (rather than exponentially) with the length m of the chain.

7. **Parameter Sharing:** Another way of reducing the number of independent parameters is through *parameter sharing*.

For example, one can assume that, in Figure 4, all of the conditional distributions

$$\mathbb{P}(X_j = \mathbf{x}_j \,|\, X_{j-1} = \mathbf{x}_{j-1}), \qquad \text{for all } j = 2, \cdots, m,$$

are governed by the *same* set of $K(K-1)$ parameters. Together with the $K-1$ parameters governing the distribution of X_1, this gives a total of

$$(K-1) + K(K-1) = K^2 - 1$$

parameters that need to be specified in order to define the joint distribution.

II.3 Linear Gaussian Models

1. **Setup and Assumptions:** Consider an arbitrary directed acyclic graph over m variables in which the j-th node represents a single continuous random variable X_j having a Gaussian distribution with

(a) the mean being a linear combination of the states of its *parent* nodes pa_j of the j-th node, and

(b) the variance being σ_j^2,

that is,

$$X_j \,|\, \mathrm{pa}_j \sim \mathrm{Normal}\Bigg(b_j + \sum_{\{\ell | X_\ell \in \mathrm{pa}_j\}} w_{j,\ell} X_\ell, \sigma_j^2\Bigg). \tag{3}$$

2. **Joint Density Function:** The logarithm of the joint density function of all nodes is

$$\begin{aligned}\sum_{j=1}^m \log\varphi\Bigg(X_j \,\Bigg|\, b_j + \sum_{\{\ell | X_\ell \in \mathrm{pa}_j\}} w_{j,\ell} X_\ell, \sigma_j^2\Bigg) \\ = -\frac{1}{2}\sum_{j=1}^m \frac{1}{\sigma_j^2}\Bigg(X_j - b_j - \sum_{\{\ell | X_\ell \in \mathrm{pa}_j\}} w_{j,\ell} X_\ell\Bigg)^2 - \frac{1}{2}\sum_{j=1}^m \log\sigma_j^2 + \mathrm{const},\end{aligned} \tag{4}$$

where "const" denotes terms that are independent of $\{b_j\}_j$, $\{w_{j,\ell}\}_{j,\ell}$ and $\{\sigma_j^2\}_j$ and $X_1, X_2, \cdots, X_m$.

In particular, note that the joint density function of all nodes is a Gaussian density function, i.e., $(X_1, X_2, \cdots, X_m)$ has a multivariate Gaussian distribution.

3. **Determining Mean and Variance of $(X_1, X_2, \cdots, X_m)$:** Assume that the nodes are numbered such that each node has a higher number than its parents. We can determine the mean and covariance of the joint distribution *recursively*.

For all $j = 1, 2, \cdots, m$, we can rewrite (3) as

$$X_j \,|\, \mathrm{pa}_j = b_j + \sum_{\{\ell | X_\ell \in \mathrm{pa}_j\}} w_{j,\ell} X_\ell + \sigma_j \varepsilon_j, \tag{5}$$

where $(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m)^\top \sim \mathrm{Normal}_m(\mathbf{0}_m, \mathbf{I}_m)$.

(a) *Determining the Mean:* Taking the expectation of (5) yields

$$\begin{aligned}\mathbb{E}[X_j] &= \mathbb{E}\big[\mathbb{E}[X_j \,|\, \mathrm{pa}_j]\big] \\ &= \mathbb{E}\Bigg[b_j + \sum_{\{\ell | X_\ell \in \mathrm{pa}_j\}} w_{j,\ell} X_\ell + \sigma_j \varepsilon_j\Bigg] \\ &= b_j + \sum_{\{\ell | X_\ell \in \mathrm{pa}_j\}} w_{j,\ell} X_\ell.\end{aligned}$$

Thus, we can determine $(\mathbb{E}[X_1], \mathbb{E}[X_2], \cdots, \mathbb{E}[X_m])^\top$ by starting at the lowest numbered node and working recursively through the graph.

(b) *Determining the Covariance:* The covariance between X_j and X_k, for all $j, k = 1, 2, \cdots, m$, is

$$\begin{aligned}\mathrm{Cov}(X_j, X_k) &= \mathbb{E}\Big[(X_j - \mathbb{E}[X_j])(X_k - \mathbb{E}[X_k])\Big] \\ &= \sum_{\{\ell | \ell \in \mathrm{pa}_j\}} \sum_{\{\ell' | \ell' \in \mathrm{pa}_k\}} w_{j,\ell} w_{k,\ell'} \,\mathrm{Cov}(X_\ell, X_{\ell'}) + \sigma_k^2 \mathbb{1}_{\{j=k\}},\end{aligned}$$

where

$$\mathbb{1}_{\{j=k\}} = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{otherwise.} \end{cases}$$

It follows that the covariance can similarly be evaluated recursively starting from the lowest numbered node.

4. **Special Case 1:** Suppose that there are *no* links in the graph so that all random variables are independent.

   (a) *Reduction of Number of Parameters:* There are no parameters $w_{j,\ell}$'s. One only needs to determine m parameters $b_1, b_2, \cdots, b_m$ for the mean and m parameters $\sigma_1^2, \sigma_2^2, \cdots, \sigma_m^2$ for the variance.

   (b) *Mean Parameters:* The mean $(\mathbb{E}[X_1], \mathbb{E}[X_2], \cdots, \mathbb{E}[X_m])^\top$ is given by

   $$(b_1, b_2, \cdots, b_m)^\top.$$

   (c) *Covariance Matrix:* The covariance matrix is diagonal of the form

   $$\operatorname{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_m^2).$$

5. **Special Case 2:** Consider a fully connected graph in which each node has all lower numbered nodes as parents.

   (a) *Total Number of Weight Parameters:* The matrix $w_{j,\ell}$ has $j-1$ entries on the j-th row and hence is a lower triangular matrix, giving a total of $m(m-1)/2$ parameters to determine.

   (b) *Total Number of Parameters in the Mean:* The parameters in the mean to be determined are $\{w_{j,\ell}\}_{j,\ell}$ and $\{b_j\}_j$, giving a total of $m(m+1)/2$ parameters to determine.

   (c) *Total Number of Parameters in the Covariance Matrix:* The parameters in the covariance matrix to be determined are $\{w_{j,\ell}\}_{j,\ell}$ and $\{\sigma_j^2\}_j$, giving a total of $M(M+1)/2$ parameters to determine.

   (d) *Summary:* In this special case, the total number of parameters to be determined is

   $$\frac{m(m-1)}{2} + m + m = \frac{m(m+3)}{2}.$$

6. **Special case 3:** Consider the intermediate level of complexity of three random variables shown in Figure 5.

   By inspecting the structure of the graph and using the results above, the mean vector is given by

   $$\begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \mathbb{E}[X_3] \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 + w_{2,1}X_1 \\ b_3 + w_{3,2}X_2 + w_{3,2}w_{2,1}X_1 \end{pmatrix},$$

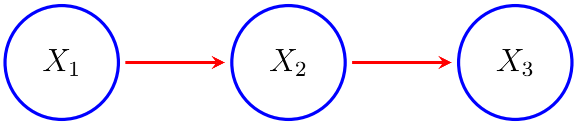

Figure 5: A special case of trivariate discrete random variable.

and the covariance matrix is given by

$$\begin{pmatrix} \sigma_1^2 & w_{2,1}\sigma_1^2 & w_{3,2}w_{2,1}\sigma_1^2 \\ w_{2,1}\sigma_1^2 & \sigma_2^2 + w_{2,1}^2\sigma_1^2 & w_{3,2}(\sigma_2^2 + w_{2,1}^2\sigma_1^2) \\ w_{3,2}w_{2,1}\sigma_1^2 & w_{3,2}(\sigma_2^2 + w_{2,1}^2\sigma_1^2) & \sigma_3^2 + w_{3,2}(\sigma_2^2 + w_{2,1}^2\sigma_1^2) \end{pmatrix}.$$

III. Conditional Independence

1. **Conditional Independence:** Consider three random variables X, Y and Z, and suppose that the conditional distribution of X, given Y and Z, does *not* depend on Y, that is,

$$f_{X\,|\,Y,Z}(x\,|\,y,z) = f_{X\,|\,Z}(x\,|\,z). \tag{6}$$

We say that X is *conditionally independent* of Y given Z, and is denote by

$$X \perp\!\!\!\perp Y\,|\,Z. \tag{7}$$

Remark. Alternatively, assuming X is conditionally independent of Y given Z, we have

$$\begin{aligned} f_{X,Y\,|\,Z}(x,y\,|\,z) &= f_{X\,|\,Y,Z}(x\,|\,y,z)f_{Y\,|\,Z}(y\,|\,z) \\ &= f_{X\,|\,Z}(x\,|\,z)f_{Y\,|\,Z}(y\,|\,z). \end{aligned}$$

Thus, conditioning on Z, the joint distribution of X and Y factorizes into the product of the marginal distribution of X and that of Y (again both conditioned on Z *only*). This says that the variables X and Y are statistically independent, given Z.

2. **Connection between Conditional Independence and Graphical Models:** An important and elegant feature of graphical models is that conditional independence properties of the joint distribution can be read directly from the graph *without* having to perform any analytical manipulations.

3. **Example 1:** Consider the example shown in Figure 6.

The corresponding joint pdf can be factorized as

$$f_{X,Y,Z}(x,y,z) = f_{X\,|\,Z}(x\,|\,z)f_{Y\,|\,Z}(y\,|\,z)f_Z(z).$$

Marginalizing with respect to Z yields

$$f_{X,Y}(x,y) = \int f_{X,Y,Z}(x,y,z)\mathrm{d}z = \int f_{X\,|\,Z}(x\,|\,z)f_{Y\,|\,Z}(y\,|\,z)f_Z(z)\mathrm{d}z.$$

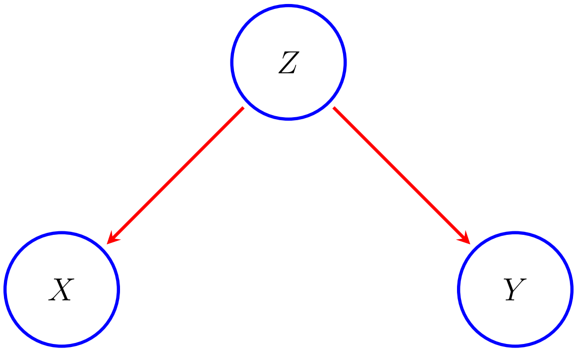

Figure 6: Example 1 of conditional independence.

In general, we cannot factorize the preceding equation into $f_X(x)f_Y(y)$ and so X and Y are *not* independent.

Now, conditioning on Z, we have

$$f_{X,Y\,|\,Z} = \frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)} = f_{X\,|\,Z}(x\,|\,z)f_{Y\,|\,Z}(y\,|\,z),$$

and we obtain $X \perp\!\!\!\perp Y \,|\, Z$.

Remark. The node Z is said to be *tail-to-tail* with respect to this path because the node corresponding to Z is connected to the tails of the two arrows, and the presence of such a path connecting nodes X and Y causes these nodes to be dependent.

4. **Example 2:** Consider the example shown in Figure 7.

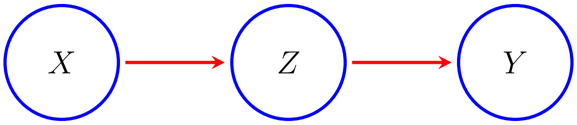

Figure 7: Example 2 of conditional independence.

The corresponding joint pdf can be factorized as

$$f_{X,Y,Z}(x,y,z) = f_X(x)f_{Z\,|\,X}(z\,|\,x)f_{Y\,|\,Z}(y\,|\,z).$$

Marginalizing with respect to Z yields

$$\begin{aligned} f_{X,Y}(x,y) = \int f_{X,Y,Z}(x,y,z)\mathrm{d}z &= f_X(x)\int f_{Z\,|\,X}(z\,|\,x)f_{Y\,|\,Z}(y\,|\,z)\mathrm{d}z \\ &= f_X(x)f_{Y\,|\,X}(y\,|\,x), \end{aligned}$$

which, in general, cannot be factorized into $f_X(x)f_Y(y)$ and so X and Y are *not* independent.

Now, conditioning on Z, we have

$$f_{X,Y\,|\,Z} = \frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)} = \frac{f_X(x)f_{Z\,|\,X}(z\,|\,x)f_{Y\,|\,Z}(y\,|\,z)}{f_Z(z)} = f_{X\,|\,Z}(x\,|\,z)f_{Y\,|\,Z}(y\,|\,z),$$

and we obtain $X \perp\!\!\!\perp Y \mid Z$.

Remark. The node Z is said to be *head-to-tail* with respect to the path from node X to node Y. Such a path connects nodes X and Y and renders them dependent.

5. **Example 3:** Consider the example shown in Figure 8.

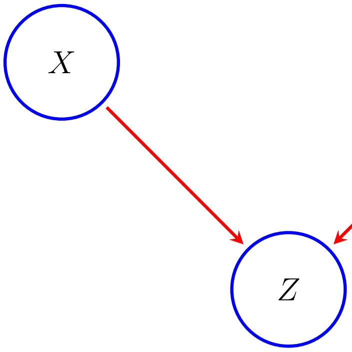

Figure 8: Example 3 of conditional independence.

The corresponding joint pdf can be factorized as

$$f_{X,Y,Z}(x,y,z) = f_X(x) f_Y(y) f_{Z|X,Y}(z \mid x,y).$$

Marginalizing with respect to Z yields

$$f_{X,Y}(x,y) = \int f_{X,Y,Z}(x,y,z)\mathrm{d}z = \int f_X(x) f_Y(y) f_{Z|X,Y}(z \mid x,y)\mathrm{d}z = f_X(x) f_Y(y),$$

which states X and Y are independent.

Now, conditioning on Z, we have

$$f_{X,Y|Z} = \frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)} = \frac{f_X(x) f_Y(y) f_{Z|X,Y}(z \mid x,y)}{f_Z(z)},$$

which, in general, cannot be factored into $f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z)$.

Remark. The node Z is *head-to-head* with respect to the path from X to Y because it connects to the heads of the two arrows. When node Z is unobserved, it "blocks" the path, and the variables X and Y are independent. However, conditioning on Z "unblocks" the path and renders X and Y dependent.

6. **Descendant:** We say that node Y is a *descendant* of node X if there is a path from X to Y in which each step of the path follows the directions of the arrows.

7. **Summary:**

   (a) A tail-to-tail node or a head-to-tail node leaves a path unblocked unless it is observed in which case it blocks the path;

   (b) A head-to-head node blocks a path if it is unobserved, but once the node, and/or at least one of its descendants, is observed, the path becomes unblocked.

8. **D-Separation:** Consider a directed acyclic graph in which A, B, and C are arbitrary non-overlapping sets of nodes (whose union may be smaller than the complete set of nodes in the graph), and consider all possible paths from any node in A to any node in B. Any such path is said to be *blocked* if it includes a node such that either

   (a) the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C, or

   (b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C.

   If *all* paths are blocked, then A is said to be *d-separated* from B by C, and the joint distribution over all of the variables in the graph satisfies $A \perp\!\!\!\perp B \,|\, C$.

IV. Undirected Graphs

IV.1 Introduction

1. **Markov Random Field:** An *undirected graph*, also knowns as a *Markov random field* or a *Markov network*, has

   (a) a set of nodes each of which corresponds to a variable or group of variables, and

   (b) a set of edges each of which connects a pair of nodes. The edges are undirected, that is, they do *not* carry arrows.

2. **Adjacent Vertices:** Two vertices X and Y are called *adjacent* if there is an edge joining them, denoted by $X \sim Y$.

3. **Path:** A *path* $X_1, X_2, \cdots, X_n$ is a set of vertices that are joined, that is, $X_{i-1} \sim X_i$ for $i = 2, \cdots, n$.

4. **Complete Graph:** A *complete graph* is a graph with every pair of vertices joined by an edge.

5. **Subgraph:** A *subgraph* $U \subset V$ is a subset of vertices together with their edges.

6. **Pairwise Markov Independence:** In a Markov random field $\mathcal{G}$, the absence of an edge implies that the corresponding random variables are *conditionally independent* given the variables at the other vertices; that is,

$$\text{no edge joining } X \text{ and } Y \iff X \perp\!\!\!\perp Y \,|\, \text{rest}, \tag{8}$$

   where "rest" refers to all of the other vertices in the graph. These are known as the *pairwise Markov independencies* of $\mathcal{G}$.

7. **Separator:** If A, B and C are subgraphs, then C is said to *separate* A and B if every path between A and B intersects a node in C. The subgraph C is said to be a *separator*.

8. **Global Markov Properties:** Separators have the property that they break the graph into conditionally independent pieces. In a Markov random field $\mathcal{G}$ with subgraphs A, B and C,

$$\text{if } C \text{ separates } A \text{ and } B, \text{ then } A \perp\!\!\!\perp B \,|\, C. \tag{9}$$

This is known as the *global Markov properties* of $\mathcal{G}$.

9. **Equivalence of Pairwise and Global Markov Property:** The pairwise and global Markov properties of a graph are equivalent (for graphs with positive distributions). The set of graphs with associated probability distributions that satisfy the pairwise Markov independencies and global Markov properties are the same.

Remark. The global Markov property allows us to decompose graphs into smaller more manageable pieces and leads to simplifications in computation and interpretation.

10. **Clique:** A *clique* is a complete subgraph, i.e., a set of vertices in a graph such that there exists an edge between all pairs of vertices in the subset. In other words, the set of vertices in a clique is fully connected.

A clique is said to be *maximal* if it is a clique and no other vertices can be added to it and still yield a clique.

Example. Consider the graph in Figure 9.

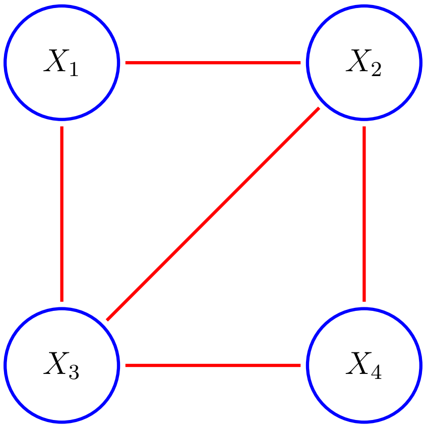

Figure 9: Clique example.

This graph has

(a) five cliques of two nodes given by $\{X_1, X_2\}$, $\{X_2, X_3\}$, $\{X_3, X_4\}$, $\{X_4, X_2\}$, and $\{X_1, X_3\}$, and

(b) two maximal cliques given by $\{X_1, X_2, X_3\}$ and $\{X_2, X_3, X_4\}$.

In particular, note that the set $\{X_1, X_2, X_3, X_4\}$ is *not* a clique.

11. **Implications on Density Factorization:** We can define the factors in the decomposition of the joint distribution of vertices in a graph to be functions of the variables in the *maximal cliques*, without loss of generality, because other cliques must be subsets of maximal cliques.

12. **Probability Density Function over a Graph:** A probability density function f over a Markov graph $\mathcal{G}$ can be can represented as

$$f(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(\mathbf{x}_C), \tag{10}$$

where $\mathcal{C}$ is the set of maximal cliques, and the positive functions $\psi_C(\cdot)$ are called the *clique potentials*. The quantity

$$Z := \int_{\mathcal{X}} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C) \mathrm{d}\mathbf{x} \tag{11}$$

is the normalizing constant, also known as the *partition function*.

Remark 1. We do *not* require the choice of ψ_C's to have a specific probabilistic interpretation as marginal or conditional distributions.

On the contrary, in directed graphs, each factor represents the conditional distribution of the corresponding variable, conditioned on the state of its parents.

Remark 2. Since the choice of $\{\psi_C\}$ is somewhat *arbitrary*, their product will in general *not* be correctly normalized. This is why we need Z in (10).

On the contrary, for directed graphs, the joint distribution was automatically normalized as a consequence of the normalization of each of the conditional distributions in the factorization.

Remark 3. The presence of Z in (10) is one of the major limitations of undirected graphs. The partition function Z is needed for parameter estimation because it will be a function of any parameters that govern the potential functions. Its computation requires integration over a high-dimensional space.

For evaluation of *local conditional distributions*, however, the partition function is *not* needed because a conditional density function is the ratio of two marginals, and the partition function cancels when evaluating this ratio.

13. **Hammersley-Clifford Theorem:** Assume that each clique potential is strictly positive, i.e., $\psi_C(\mathbf{x}_C) > 0$ for all possible values of $\mathbf{x}_C$. Let

    (a) S_1 be the set of such distributions that are consistent with the set of conditional independence statements that can be read from the graph using graph separation, and

    (b) S_2 be the set of such distributions that can be expressed as a factorization of the form (10) with respect to the maximal cliques of the graph.

Then, the sets S_1 and S_2 are identical.

Remark. Hammersley-Clifford theorem made the formal connection between conditional independence and factorization for undirected graphs.

14. **Energy Function:** Because we are restricted to potential functions which are *strictly positive*, it is convenient to express them as exponentials, so that

$$\psi_C(\mathbf{x}_C) = \exp\big(-E(\mathbf{x}_C)\big), \tag{12}$$

where E is called an *energy function*, and the exponential representation is called the *Boltzmann distribution*.

Remark. The joint distribution is defined as the product of potentials, and so the total energy is obtained by adding the energies of each of the maximal cliques.

IV.2 Relationships between Directed and Undirected Graphs

1. **Moralization:** In order to convert a directed graph into an undirected graph and convert any distribution specified by a factorization over a directed graph into one specified by a factorization over an undirected graph, we do the following:

   (a) Add additional undirected links between *all* pairs of parents for each node in the graph;

   (b) Drop the arrows on the original links;

   (c) Initialize all of the clique potentials of the moral graph to 1;

   (d) Take each conditional distribution factor in the original directed graph and multiply it into one of the clique potentials.

   This process is called *moralization*, and the resulting undirected graph is called a *moral graph*.

   Remark 1. There will always exist at least one maximal clique that contains all of the variables in the factor as a result of the moralization step.

   Remark 2. In all cases, the partition function is given by $Z = 1$.

   Remark 3. When converting a directed graph into an undirected one, we could always *trivially* convert any distribution over a directed graph into one over an undirected graph by simply using a fully connected undirected graph. This would, however, discard *all* conditional independence properties and so would be useless. The process of moralization adds the *fewest* extra edges and so retains the maximum number of independence properties.

2. **Example:** Consider the directed graph in the left panel in Figure 10. The joint distribution of the direct graph is

$$f_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4) = f_{X_1}(x_1)f_{X_2}(x_2)f_{X_3}(x_3)f_{X_4|X_1,X_2,X_3}(x_4 \,|\, x_1,x_2,x_3).$$

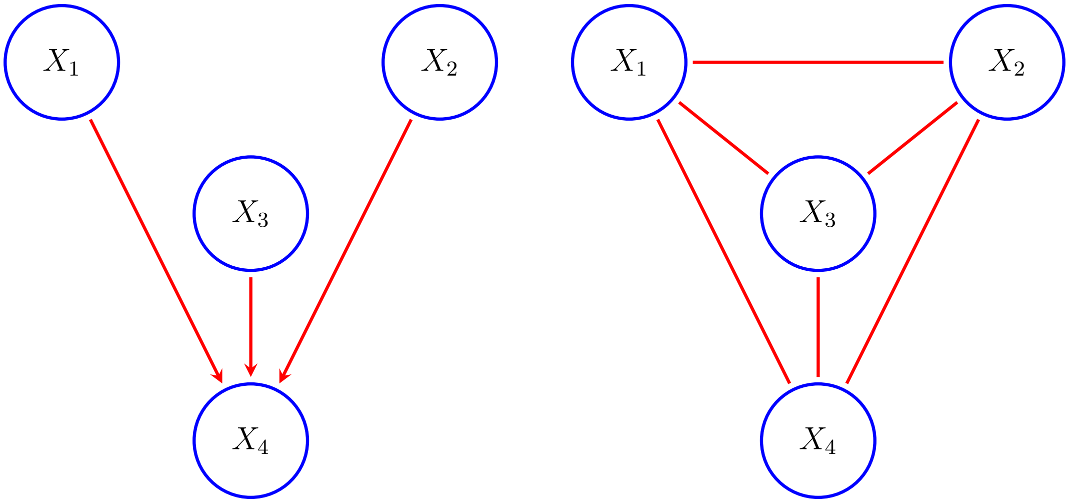

Figure 10: Moralization example.

Note that the last factor $f_{X_4 | X_1, X_2, X_3}(x_4 \,|\, x_1, x_2, x_3)$ involves all four variables, and, hence, they all must belong to the single clique if this conditional distribution is absorbed into a clique potential. Therefore, we add extra edges between all pairs of parents of the vertex X_4. The resulting moral graph is shown in the right panel of Figure 10 and the clique potential is

$$\psi_C(x_1, x_2, x_3, x_4) = f_{X_1, X_2, X_3, X_4}(x_1, x_2, x_3, x_4).$$

Remark. In this particular example, the moral graph is fully connected and so exhibits *no* conditional independence properties, in contrast to the original directed graph.

V. Undirected Graphical Models for Continuous Random Variables

V.1 Introduction

1. **Setup:** In this section, we consider the undirected graphs where all the variables are continuous.

2. **Assumption:** Assume that the observations have a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

   Remarks.

   (a) Since the Gaussian distribution represents at most *second-order* relationships, it automatically encodes a *pairwise Markov graph*.

   (b) The Gaussian distribution has the property that all conditional distributions are also Gaussian.

3. **Inverse Covariance Matrix of a Gaussian Distribution and Conditional Independence:** Suppose

$$X = (X_1, X_2, \cdots, X_p)^\top \sim \mathrm{Normal}_p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is positive definite. If (i,j)-th entry of $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$ is zero, then X_i and X_j are conditionally independent, given all other variables.

Remark. This result can be proved by working with the inverse of a block matrix and the conditional distribution of a Gaussian random vector.

4. **Conditional Gaussian Distribution:** Suppose we partition a p-dimensional random vector X as $X = (Z^\top, Y^\top)^\top$, where $Z = (X_1, X_2, \cdots, X_{p-1})^\top \in \mathbb{R}^{p-1}$ consists of the first $(p-1)$ components of X and $Y = X_p$ is the last. Then, the conditional distribution of Y given $Z = \mathbf{z}$ is

$$Y \,|\, Z = \mathbf{z} \sim \mathrm{Normal}\big(\mu_Y + (\mathbf{z} - \boldsymbol{\mu}_Z)^\top \boldsymbol{\Sigma}_{ZZ}^{-1} \boldsymbol{\Sigma}_{ZY}, \sigma_{YY} - \boldsymbol{\Sigma}_{YZ} \boldsymbol{\Sigma}_{ZZ}^{-1} \boldsymbol{\Sigma}_{ZY}\big), \tag{13}$$

where the covariance matrix is partitioned as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{ZZ} & \boldsymbol{\Sigma}_{ZY} \\ \boldsymbol{\Sigma}_{YZ} & \sigma_{YY} \end{pmatrix}. \tag{14}$$

Notice that the conditional mean in (13) has exactly the same form as the population multiple linear regression of Y on Z, with the regression coefficient $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{ZZ}^{-1} \boldsymbol{\Sigma}_{ZY}$.

If we partition the inverse covariance matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ in a similar fashion, we have

$$\boldsymbol{\Theta}_{ZY} = -\theta_{YY} \boldsymbol{\Sigma}_{ZZ}^{-1} \boldsymbol{\Sigma}_{ZY}, \tag{15}$$

where $\theta_{YY}^{-1} = \sigma_{YY} - \boldsymbol{\Sigma}_{YZ} \boldsymbol{\Sigma}_{ZZ}^{-1} \boldsymbol{\Sigma}_{ZY} > 0$. It follows that

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{ZZ}^{-1} \boldsymbol{\Sigma}_{ZY} = -\frac{\boldsymbol{\Theta}_{ZY}}{\theta_{YY}}. \tag{16}$$

5. **$\boldsymbol{\Theta}$ as the Natural Parameter:** The inverse covariance matrix $\boldsymbol{\Theta}$ captures *all* the second-order structural and quantitative information to describe the conditional distribution of each node given the rest.

The distribution arising from a Gaussian graphical model is a *Wishart distribution*, and is a member of the exponential family, with canonical or "natural" parameter being $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$.

V.2 Estimation of the Parameters when the Graph Structure is Known

1. **Complete Graph Case:**

   (a) *Assumptions:* We assume the graph is complete, i.e., the graph is fully connected.

(b) *Data:* Suppose we have n multivariate normal realizations $\mathbf{x}_i \in \mathbb{R}^p$, for $i = 1, 2, \cdots, n$, with population mean and covariance $\mathbf{\Sigma}$.

(c) *Sample Covariance Matrix:* Let $\mathbf{S} \in \mathbb{R}^{p\times p}$ be the empirical covariance matrix, where

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top, \tag{17}$$

and $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ is the sample mean vector.

(d) *Log-likelihood Function and MLE of $\mathbf{\Theta}$:* The log-likelihood function of $\mathbf{\Theta}$ is

$$\ell(\mathbf{\Theta}) = \log\det(\mathbf{\Theta}) - \mathrm{trace}(\mathbf{S\Theta}). \tag{18}$$

The quantity $-\ell$ is a convex function of $\mathbf{\Theta}$, and it is easy to show the maximum likelihood estimate of $\mathbf{\Theta}$ is $\mathbf{S}^{-1}$. It follows that the maximum likelihood estimate of $\mathbf{\Sigma}$ is $\mathbf{S}$.

2. Incomplete Graph with Known Structure:

(a) *Setup:* In the incomplete graph case, some of the edges are missing, implying that the corresponding entries of $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ are 0 in the Gaussian distribution case.

(b) *General Approach:* We maximize the likelihood function under the constraints that some pre-defined subset of the parameters are zero, which is an equality-constrained convex optimization problem.

Remark. The central idea of the approach described above is to decompose the graph into its *maximal cliques*.

(c) *Optimization Problem:* We maximize the following constrained log-likelihood function

$$\ell_C(\mathbf{\Theta}) = \log\det(\mathbf{\Theta}) - \mathrm{trace}(\mathbf{S\Theta}) - \sum_{(j,k)\notin E}\gamma_{j,k}\theta_{j,k}, \tag{19}$$

where the last term $\sum_{(j,k)\notin E}\gamma_{j,k}\theta_{j,k}$ restricts the (j,k)-th entry of $\mathbf{\Theta}$ to be zero and reflects the known graph structure there is no edge between the corresponding variables.

The first-order optimality condition of maximizing ℓ_C requires

$$\mathbf{\Theta}^{-1} - \mathbf{S} - \mathbf{\Gamma} = \mathbf{0}_{p\times p}, \tag{20}$$

where $\mathbf{\Gamma}$ is a matrix of Lagrange parameters with nonzero values for all pairs with edges absent. In deriving (20), we use the fact that the derivative of $\log\det(\mathbf{\Theta})$ is $\mathbf{\Theta}^{-1}$.

(d) *Characterizing the Solution:* Let $\mathbf{W} = \mathbf{\Theta}^{-1}$. Without the loss of generality, we partition the matrix $\mathbf{W} \in \mathbb{R}^{p\times p}$ as

$$\mathbf{W} = \begin{pmatrix}\mathbf{W}_{11} & w_{12}\\ w_{21} & w_{22}\end{pmatrix}$$

and look at the last column and the last row of $\mathbf{W}$. We have

$$\mathbf{W}\boldsymbol{\Theta} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & w_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\Theta}_{12} \\ \boldsymbol{\Theta}_{21} & \theta_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{(p-1)\times(p-1)} & \mathbf{0}_{p-1} \\ \mathbf{0}_{p-1}^\top & 1 \end{pmatrix}. \tag{21}$$

Therefore, we must have

$$\mathbf{W}_{11}\boldsymbol{\Theta}_{12} + \mathbf{W}_{12}\theta_{22} = 0,$$

from which we obtain

$$\mathbf{W}_{12} = -\mathbf{W}_{11}\frac{\boldsymbol{\Theta}_{12}}{\theta_{22}} = \mathbf{W}_{11}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = -\frac{\boldsymbol{\Theta}_{12}}{\theta_{22}} \in \mathbb{R}^{p-1}$. From (20), using a similar partition of the matrices $\mathbf{S}$ and $\boldsymbol{\Gamma}$, it follows that

$$\mathbf{0}_{p-1} = \mathbf{W}_{12} - \mathbf{S}_{12} - \boldsymbol{\Gamma}_{12} = \mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{S}_{12} - \boldsymbol{\Gamma}_{12}. \tag{22}$$

To solve this equation, suppose that there are $p-q-1$ non-zero elements in $\boldsymbol{\Gamma}_{12}$, i.e., $p-q-1$ edges are constrained to be zero. Then, these $p-q-1$ rows carry no information and can be removed. Consequently, we can reduce $\boldsymbol{\beta}$ to $\widetilde{\boldsymbol{\beta}}$ by removing the corresponding $p-q-1$ zero elements and obtain the following reduced $q \times q$ system of equations

$$\widetilde{\mathbf{W}}_{11}\widetilde{\boldsymbol{\beta}} - \widetilde{\mathbf{S}}_{12} = \mathbf{0}_q, \tag{23}$$

from which we have $\widetilde{\boldsymbol{\beta}} = \widetilde{\mathbf{W}}_{11}^{-1}\widetilde{\mathbf{S}}_{12}$. To obtain $\boldsymbol{\beta}$ that satisfies (22), simply insert zeros at appropriate positions.

It is easy to obtain from (21) that

$$\mathbf{W}_{12}\boldsymbol{\Theta}_{12} + w_{22}\theta_{22} = 1,$$

from which we can solve

$$\frac{1}{\theta_{22}} = w_{22} - \mathbf{W}_{12}^\top\boldsymbol{\beta}.$$

Finally, we have $w_{22} = s_{22}$, since the diagonal of $\boldsymbol{\Gamma}$ is zero.

(e) *Algorithm:* The arguments above lead to a simple iterative procedure for estimating both $\mathbf{W}$ and its inverse $\boldsymbol{\Theta}$, subject to the constraints of the missing edges. The complete iterative algorithm is shown in Algorithm 1.

Algorithm 1 A Modified Regression Algorithm for Estimation of an Undirected Gaussian Graphical Model with Known Structure

```
Initialize $\mathbf{W} = \mathbf{S}$;
repeat
    for $j = 1, 2, \cdots, p$ do
        Partition the matrix $\mathbf{W}$ into one part containing all but the $j$-th row and
        column and the other part containing the $j$-th row and column;
        Solve
            $$\widetilde{\mathbf{W}}_{11}\widetilde{\boldsymbol{\beta}} - \widetilde{\mathbf{S}}_{12} = \mathbf{0}$$
        for the unconstrained edge parameters $\widetilde{\boldsymbol{\beta}}$, using the reduced system of
        equations as in (23);
        Obtain $\boldsymbol{\beta}$ that satisfies (22) by padding $\widetilde{\boldsymbol{\beta}}$ with zeros in the appropriate
        positions;
        Update $\mathbf{W}_{12} = \mathbf{W}_{11}\boldsymbol{\beta}$;
    end for
until convergence
In the final cycle for each $j$, solve for
    $$\boldsymbol{\Theta}_{12} = -\boldsymbol{\beta}\theta_{22}, \qquad w_{22} = s_{22}, \qquad \text{and} \qquad \theta_{22}^{-1} = w_{22} - \mathbf{W}_{12}^{\top}\boldsymbol{\beta}.$$
```

V.3 Estimation of Graph Structure

1. **Motivation:** In most cases of applications, we do *not* know the exact graph structure or the missing edges, and would like to try to discover from the data.

2. **General Idea:** The general idea is that, rather than estimating the full covariance $\boldsymbol{\Sigma}$ or its inverse $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, only estimate which components of $\theta_{i,j}$ are non-zero, by fitting a lasso regression using each variance as the response and the others as predictors.

3. **Lasso-Type Optimization Problem:** The idea above leads to the problem of maximizing the penalized log-likelihood function

$$\ell_{\text{lasso}}(\boldsymbol{\Theta}) := \log\det(\boldsymbol{\Theta}) - \operatorname{trace}(\mathbf{S}\boldsymbol{\Theta}) - \lambda \cdot \|\boldsymbol{\Theta}\|_1, \tag{24}$$

where $\|\boldsymbol{\Theta}\|_1$ is the L^1-norm, the sum of the absolute values of the elements in $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. Notice that the negative of (24) is a convex function of $\boldsymbol{\Theta}$.

The gradient of ℓ_{lasso} is

$$\boldsymbol{\Theta}^{-1} - \mathbf{S} - \lambda \cdot \operatorname{sign}(\boldsymbol{\Theta}) = \mathbf{0}_{p\times p},$$

where the sign function is applied entry-wise with

$$\operatorname{sign}(\theta) = \begin{cases} -1, & \text{if } \theta < 0, \\ [-1, 1], & \text{if } \theta = 0, \\ +1, & \text{if } \theta > 0. \end{cases}$$

By letting $\mathbf{W} = \mathbf{\Theta}^{-1}$ and partitioning $\mathbf{W}$ in a similar fashion to the previous section, we have

$$\mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{S}_{12} + \lambda \cdot \operatorname{sign}(\boldsymbol{\beta}) = \mathbf{0}_{p-1}, \tag{25}$$

where we flip the sign in the last term since $\boldsymbol{\beta} = -\frac{\mathbf{\Theta}_{12}}{\theta_{22}}$ and $\boldsymbol{\beta}$ and $\mathbf{\Theta}_{12}$ have opposite signs.

To proceed, we use the coordinate descent method at each stage. Let $\mathbf{V} = \mathbf{W}_{11}$, the update has the form

$$\beta_j \quad \longleftarrow \quad S_\lambda \left(s_{12,j} - \sum_{k \neq j} v_{k,j} \beta_k \right) / V_{jj}, \tag{26}$$

for $j = 1, 2, \cdots, p-1$, where $s_{12,j}$ is the j-th element of $\mathbf{S}_{12}$, $v_{k,j}$ is the (k, j)-th entry of $\mathbf{V}$, and S_λ is the soft-thresholding operator

$$S_\lambda(x) = \operatorname{sign}(x) \cdot (|x| - \lambda)_+.$$

The diagonal elements w_{jj} of the matrix $\mathbf{W}$ are simply $s_{jj} + \lambda$ and are fixed.

4. **Graphical Lasso Algorithm:** The entire algorithm of solving $\mathbf{\Theta}$ is called *graphical lasso*, summarized in Algorithm 2.

Algorithm 2 Graphical Lasso

```
Initialize $\mathbf{W} = \mathbf{S} + \lambda \cdot \mathbf{I}_p$. The diagonal of $\mathbf{W}$ remains unchanged in the following steps;
repeat
    for $j = 1, 2, \cdots, p$ do
        Partition the matrix $\mathbf{W}$ into one part containing all but the $j$th row and column and the other part containing the $j$th row and column;
        Solve the estimated equations
            $\mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{S}_{12} + \lambda \cdot \operatorname{sign}(\boldsymbol{\beta}) = \mathbf{0}_{p-1};$
        Update $\mathbf{W}_{12} = \mathbf{W}_{11}\boldsymbol{\beta}$;
    end for
until convergence
In the final cycle for each $j$, solve for
    $\theta_{22}^{-1} = s_{22} - \mathbf{W}_{12}^\top \beta \qquad \text{and} \qquad \mathbf{\Theta}_{12} = -\boldsymbol{\beta}\theta_{22}.$
```

VI. Undirected Graphical Models for Discrete Variables

1. **General Introductions:**

   (a) The pairwise Markov networks with binary variables are called the *Ising models* in the statistical mechanics literature, and the *Boltzmann machines* in the machine learning literature.

   (b) The values at each node can be observed ("visible") or unobserved ("hidden"). Nodes are organized in layers, similar to neural networks.

2. **Ising Model 1 — Introduction:** Consider all p nodes in a graph $\mathcal{G} = (V, E)$ are visible with edges pairs (j, k) enumerated in E, and let the binary-valued variable at node j be X_j. The *Ising model* for their joint probabilities is given by

$$\mathbb{P}_{\mathbf{\Theta}}(X = \mathbf{x}) = \exp\left[\sum_{(j,k)\in E} \theta_{j,k} x_j x_k - \Phi(\mathbf{\Theta})\right], \tag{27}$$

   for $\mathbf{x} = (x_1, x_2, \cdots, x_p)^\top \in \mathcal{X} := \{0, 1\}^p$. Here, Φ is the log-partition function defined as

$$\Phi(\mathbf{\Theta}) = \log\left[\sum_{\mathbf{x}\in\mathcal{X}} \exp\left(\sum_{(j,k)\in E} \theta_{j,k} x_j x_k\right)\right], \tag{28}$$

   ensuring the probability mass function add to one over $\mathcal{X}$.

   Remark 1. We require a *constant node* $X_0 = 1$ to be included with "edges" to all the other variables.

   Remark 2. In statistical literature, the Ising model is equivalent to a first-order-interaction Poisson log-linear model for multiway tables of counts.

3. **Ising Model 2 — Conditional Probability Mass Function:** For each node conditional on others, the corresponding conditional probability mass function takes on the form

$$\mathbb{P}_{\mathbf{\Theta}}(X_j = 1 \,|\, X_{-j} = \mathbf{x}_{-j}) = \frac{1}{1 + \exp(-\theta_{j,0} - \sum_{(j,k)\in E} \theta_{j,k} x_k)}, \tag{29}$$

   where X_{-j} denotes all of the nodes except X_j and $\mathbf{x}_{-j}$ denotes the observed value of X_{-j}. Here, the parameter $\theta_{j,k}$ measures the dependence of X_j on X_k, conditional on other nodes.

4. **Ising Model 3 — Estimation of Parameters When Graph Structure is Known:** Suppose we have the observations $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,p})^\top \in \mathcal{X}$ for $i = 1, 2, \cdots, n$. The average log-likelihood is

$$\ell(\mathbf{\Theta}) = \frac{1}{n}\sum_{i=1}^{n} \log \mathbb{P}_{\mathbf{\Theta}}(X = \mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{(j,k)\in E} \theta_{j,k} x_{i,j} x_{i,k} - \Phi(\mathbf{\Theta})\right]. \tag{30}$$

The gradient of the log-likelihood is

$$\frac{\partial \ell(\boldsymbol{\Theta})}{\partial \theta_{j,k}} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}x_{i,k} - \frac{\partial \Phi(\boldsymbol{\Theta})}{\partial \theta_{j,k}},$$

and

$$\frac{\partial \Phi(\boldsymbol{\Theta})}{\partial \theta_{j,k}} = \sum_{\mathbf{x}\in\mathcal{X}} x_j x_k \mathbb{P}_{\boldsymbol{\Theta}}(X = \mathbf{x}) = \mathbb{E}_{\boldsymbol{\Theta}}[X_j X_k].$$

Setting the gradient to 0, we obtain

$$\widehat{\mathbb{E}}[X_j X_k] - \mathbb{E}_{\widehat{\boldsymbol{\Theta}}}[X_j X_k] = 0, \tag{31}$$

where $\widehat{\boldsymbol{\Theta}}$ is the maximum likelihood estimate of $\boldsymbol{\Theta}$ and

$$\widehat{\mathbb{E}}[X_j X_k] = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}x_{i,k},$$

is the expectation taken with respect to the empirical distribution of the data. The central idea of (31) is that the maximum likelihood estimates match the estimated inner products between the nodes to their observed inner products.

Computation of $\widehat{\boldsymbol{\Theta}}$*:* To obtain $\widehat{\boldsymbol{\Theta}}$, we can use the gradient descent algorithm or Newton's method. However, the computation of $\mathbb{E}_{\boldsymbol{\Theta}}[X_j X_k]$ involves enumeration of $\mathbb{P}_{\boldsymbol{\Theta}}(X)$ over 2^{p-2} of the $|\mathcal{X}| = 2^p$ possible values of X, which is infeasible for large p.

5. **Hidden Nodes:** Suppose that a subset of variables X_{mis} are unobserved or "hidden", and the remainder X_{obs} are observed or "visible". The log-likelihood of the observed data is

$$\begin{aligned}\ell(\boldsymbol{\Theta}) &= \sum_{i=1}^{n} \log \mathbb{P}_{\boldsymbol{\Theta}}(X_{\text{obs}} = x_{i,\text{obs}}) \\ &= \sum_{i=1}^{n} \log\left[\sum_{\mathbf{x}_{\text{mis}}\in\mathcal{X}_{\text{mis}}} \exp\left(\sum_{(j,k)\in E} (\theta_{j,k}x_{i,j}x_{i,k} - \Phi(\boldsymbol{\Theta}))\right)\right],\end{aligned}$$

where the sum over $\mathcal{X}_{\text{mis}}$ means that we are summing over all possible $\{0,1\}$ values for the hidden units.

The gradient of ℓ with respect to $\theta_{j,k}$ is

$$\frac{\partial \ell(\boldsymbol{\Theta})}{\partial \theta_{j,k}} = \widehat{\mathbb{E}}_{\text{obs}}\big[\mathbb{E}_{\boldsymbol{\Theta}}[X_j X_k \,|\, X_{\text{obs}}]\big] - \mathbb{E}_{\boldsymbol{\Theta}}[X_j X_k], \tag{32}$$

where

(a) the first term is an empirical average of $X_j X_k$ if both are visible; if one or both are hidden, they are first imputed given the visible data, and then averaged over the hidden variables;

(b) the second term is the unconditional expectation of $X_j X_k$.

The inner expectation in the first term can be evaluated in the following way: for observation i,

(a) if both j-th and k-th components are observed, then

$$\mathbb{E}_{\boldsymbol{\Theta}}[X_j X_k \,|\, X_{\text{obs}} = \mathbf{x}_{i,\text{obs}}] = x_{i,j} x_{i,k};$$

(b) if the j-th component is observed but the k-th component is hidden, then

$$\mathbb{E}_{\boldsymbol{\Theta}}[X_j X_k \,|\, X_{\text{obs}} = \mathbf{x}_{i,\text{obs}}] = x_{i,j} \mathbb{P}_{\boldsymbol{\Theta}}(X_k = 1 \,|\, X_{\text{obs}} = \mathbf{x}_{i,\text{obs}});$$

(c) if both j-th and k-th components are hidden, then

$$\mathbb{E}_{\boldsymbol{\Theta}}[X_j X_k \,|\, X_{\text{obs}} = \mathbf{x}_{i,\text{obs}}] = \mathbb{P}_{\boldsymbol{\Theta}}(X_j = 1, X_k = 1 \,|\, X_{\text{obs}} = \mathbf{x}_{i,\text{obs}}).$$

References

Bishop, Christopher M (Aug. 2016). *Pattern Recognition and Machine Learning*. en. Springer New York.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.