STAT 516: Stochastic Modeling of Scientific Data **Autumn 2018**

Lecture 2: A Brief Introduction to Graphical Model

Instructor: Yen-Chi Chen

These notes are partially based on those of Mathias Drton.

2.1 Conditional Independence

2.1.1 Independence Revisited

Recall that two random variables X and Y are independent if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y).$$

In this case, we write it as $X \perp\!\!\!\perp Y$. Let p_X and p_Y denote the PDF or PMF of X and Y, respectively. Then independence also implies

$$p_{XY}(x, y) = p_X(x)p_Y(y) \Leftrightarrow p_{X|Y}(x|y) = p_X(x).$$

Consider a special case where both X and Y are categorical variables such that $X \in \{1, 2, \cdots, m\}$ and $Y \in \{1, 2, \cdots, n\}$. We further define

$$q_{ij} = P(X = i, Y = j) \quad q_{i+} = P(X = i) \quad q_{+j} = P(Y = j).$$

Then $X \perp\!\!\!\perp Y$ if and only if

$$q_{ij} = q_{i+} \cdot q_{+j} \quad \text{for all } i, j.$$

Lemma 2.1 *Let Q be an $m \times n$ matrix such that $Q_{ij} = q_{ij}$. Then $X \perp\!\!\!\perp Y$ if and only if the matrix Q has rank* 1*.*

Proof:

$\Rightarrow$:

This direction is easy to see because $q_{ij} = q_{i+} \cdot q_{+j}$ implies that $Q = uv^T$, where $u = (q_{1+}, q_{2+}, \cdots, q_{n+})$ and $v = (q_{+1}, q_{+2}, \cdots, q_{+m})$.

$\Leftarrow$:

If Q has rank 1, there exists vectors $u \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ such that $Q = uv^T$. Because $q_{ij} \geq 0$, we may choose every elements of u and v to be non-negative, i.e., $u_j \geq 0$ and $v_j \geq 0$ for every i and j.

Since $Q_{ij} = p_{ij} = u_i v_j$,

$$p_{i+} = \sum_{j=1,\cdots,m} p_{ij} = \sum_{j=1}^{m} u_i v_j = u_i v_+,$$

where $v_+ = \sum_{j=1}^m v_j > 0$. Similarly,

$$p_{+j} = u_+ v_j, \quad u_+ = \sum_{i=1}^n u_i.$$

Therefore, we obtain

$$u_i = \frac{p_{i+}}{v_+}, \quad v_j = \frac{p_{+j}}{u_+}$$

and

$$p_{ij} = u_i v_j = \frac{p_{i+} p_{+j}}{v_+ u_+} = p_{i+} p_{+j}$$

because $v_+ u_+ = \sum_{j=1}^m v_j \sum_{i=1}^n u_i = \sum_{i,j} u_i v_j = \sum_{i,j} p_{ij} = 1$.

■

2.1.2 Conditional Independence

For three RVs X, Y, and Z, we say X, Y are conditional independent given Z if

$$P(X \leq x, Y \leq y | Z = z) = P(X \leq x | Z = z) P(Y \leq y | Z = z)$$

for every x and y and P_Z-almost everywhere of z. P_Z-almost everywhere of z means that the above equality holds for all z except for a set of values that has 0 probability. It is a slightly weaker notion than 'for every z'. We use the notation

$$X \perp\!\!\!\perp Y | Z$$

for denote the case where X, Y are conditional independent given Z.

Note that $X \perp\!\!\!\perp Y | Z$ also implies

$$P(X \leq x | Y = y, Z = z) = P(X \leq x | Z = z)$$

for every x and $P_{Y,Z}$-almost everywhere of (y, z).

Theorem 2.2 *Let p_{XYZ} be the joint PDF/PMF of X, Y, and Z. Then the followings are equivalent:*

(i) $X \perp\!\!\!\perp Y | Z$.

(ii) $p_{XY|Z}(x, y|z) = p_{X|Z}(x|z) p_{Y|Z}(y|z)$ *a.e.*

(iii) $p_{X|YZ}(x|y, z) = p_{X|Z}(x|z)$ *a.e.*

(iv) $p_{XYZ}(x, y, z) = \frac{p_{XZ}(x,z) p_{YZ}(y,z)}{p_Z(z)}$ *a.e.*

(v) $p_{XYZ}(x, y, z) = g(x, z) h(y, z)$, *where g and h are some (measurable) functions.*

(vi) $p_{X|YZ}(x|y, z) = w(x, z)$, *where w is some (measurable) function.*

Proof: The equivalence between (i), (ii), (iii), and (iv) are trivial so we focus on case (v) and (vi).

(ii) $\Rightarrow$ (v):
Because

$$p_{XY|Z}(x, y|z) = p_{X|Z}(x|z) p_{Y|Z}(y|z),$$

we have

$$\frac{p_{XYZ}(x,y,z)}{p_Z(z)} = \frac{p_{XZ}(x,z)}{p_Z(z)}\frac{p_{YZ}(y,z)}{p_Z(z)}$$

so

$$p_{XYZ}(x,y,z) = \frac{p_{XZ}(x,z)p_{YZ}(y,z)}{p_Z(z)} = h(x,z)g(y,z),$$

which proves (v).

(v) $\Rightarrow$ (vi):
Based on (v), we have

$$p_{YZ}(y,z) = \int p_{XYZ}(x,y,z)dx = h(y,z)\int g(x,z)dx = h(y,z)q(z).$$

Thus,

$$p_{X|YZ}(x|y,z) = \frac{p_{XYZ}(x,y,z)}{p_{YZ}(y,z)} = \frac{g(x,z)h(y,z)}{h(y,z)q(z)} = \frac{g(x,z)}{q(z)} = w(x,z).$$

Finally, we show that (vi) $\Rightarrow$ (iii):

$$\begin{aligned} p_{X|Z}(x|z) = \int p_{XY|Z}(x,y|z)dy &= \int p_{X|YZ}(x|y,z)p_{Y|Z}(y|z)dy \\ &= w(x,z)\int p_{Y|Z}(y|z)dy = w(x,z) = p_{X|YZ}(x|y,z). \end{aligned}$$

■

To see the power of the above theorem, we now consider the problem of a Gaussian random vector $X = (X_1, X_2, \cdots, X_p) \in \mathbb{R}^p$ with a mean vector μ and a covariance matrix Σ. Assume that Σ is positive definite, then the joint PDF can be written as

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^p \mathsf{det}(\Sigma)}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\},$$

where $x = (x_1, \cdots, x_p)$.

In this model, there are two parameters μ and Σ. What does the conditional independence $X_1 \perp\!\!\!\perp X_2 | X_3, \cdots, X_p$ tell us about the underlying parameters?

Applying the property (v) in the above theorem, we can factorize p_X into

$$p_X(x) = g(x_1, x_3, x_4, \cdots, x_p)h(x_2, x_3, \cdots, x_p).$$

Therefore,

$$\log p_X(x) = \tilde{g}(x_1, x_3, x_4, \cdots, x_p) + \tilde{h}(x_2, x_3, \cdots, x_p) = -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) + C_0,$$

where C_0 is a constant with respect to x.

Using the fact that

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \sum_{i,j=1}^{p}(x_i-\mu_i)(x_j-\mu_j)\left(\Sigma^{-1}\right)_{ij},$$

we conclude that $\left(\Sigma^{-1}\right)_{12} = 0$. Namely, for a Gaussian random vector, if we see the (i, j)-th element of the inverse covariance matrix (also known as the precision matrix) is 0, we have the conditional independence of X_i and X_j given the other elements.

Here are five important properties of conditional independence. Let X, Y, Z, W be RVs.

(C1) (symmetry) $X \perp\!\!\!\perp Y|Z \Longleftrightarrow Y \perp\!\!\!\perp X|Z$.

(C2) (decomposition) $X \perp\!\!\!\perp Y|Z \Longrightarrow h(X) \perp\!\!\!\perp Y|Z$ for any (measurable) function h.
A special case is: $(X, W) \perp\!\!\!\perp Y|Z \Longrightarrow X \perp\!\!\!\perp Y|Z$.

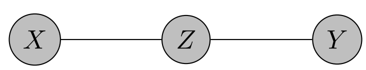
(C3) (weak union) $X \perp\!\!\!\perp Y|Z \Longrightarrow X \perp\!\!\!\perp Y|Z, h(X)$ for any (measurable) function h.
A special case is: $(X, W) \perp\!\!\!\perp Y|Z \Longrightarrow X \perp\!\!\!\perp Y|(Z, W)$

(C4) (contraction)
$$X \perp\!\!\!\perp Y|Z \text{ and } X \perp\!\!\!\perp W|(Y, Z) \Longleftrightarrow X \perp\!\!\!\perp (W, Y)|Z.$$

(C5) If the joint PDF $p_{XYZW}(x, y, z, w)$ satisfies $f_{YZW}(y, z, w) > 0$ almost everywhere. Then
$$X \perp\!\!\!\perp Y|(W, Z) \text{ and } X \perp\!\!\!\perp W|(Y, Z) \Longleftrightarrow X \perp\!\!\!\perp (W, Y)|Z.$$

2.2 Graphical Model

The conditional independence can be represented using a graph. Suppose that $X \perp\!\!\!\perp Y|Z$ so by (v) of Theorem 2.2,
$$p_{XYZ}(x, y, z) = g(x, z)h(y, z)$$
for some functions g and h. We then use the following graph to represent it their relation:

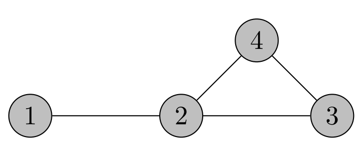


The edge $X - Z$ is drawn because the density factorization has a factor, namely $g(x, z)$, that depends on both x and z. Similarly, the edge $Z - Y$ is drawn because of factor $h(y, z)$.

Note that there is no edge between $X - Y$. The only path from X to Y passes through Z. Later we will see that in the graphical model, this implies conditional independence of X and Y given Z.

Formally, a *graph* $G = (V, E)$ is a pair consisting of a (finite) vertex set V and an edge set $E \subset V \times V$. Here, we consider *undirected graphs* where an edge $v - w$ is represented by the fact that (v, w) and (w, v) are both in E. We assume no self-loops, so $(v, v) \notin E$ for all $v \in V$.

Example 1: If $V = \{1, 2, 3, 4\}$ and
$$E = \{(1, 2), (2, 1), (2, 3), (3, 2), (2, 4), (4, 2), (3, 4), (4, 3)\}$$
then the picture is



A non-empty subset of nodes $A \subseteq V$ is *complete* if there is an edge $v - w$ between any pair of nodes $v, w \in A$. Complete sets are also called *cliques*. Sometimes, clique refers to an inclusion-maximal complete set. We denote the family of all complete sets as $\mathcal{C}(G)$.

In the above example, complete sets are

$$\{1\},\ \{2\},\ \{3\},\ \{4\}, \{1,2\},\ \{2,3\},\ \{2,4\},\ \{3,4\}, \{2,3,4\}.$$

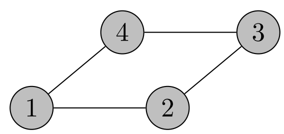
And inclusion maximal complete sets are $\{1,2\}$, $\{2,3,4\}$.

A *graphical model* uses a graph to represent the conditional independence between a set of RVs. Let $G = (V, E)$ be a graph. Let $X = (X_v : v \in V)$ be a random vector, with coordinates indexed by the nodes of G. The distribution of X is said to *factorize according to* G if it has a density $p_X(x)$ such that

$$p_X(x) = \prod_{C \in \mathcal{C}(G)} \psi_C(x_v : v \in C), \quad x \in \mathbb{R}^V.$$

Here, $\psi_C : \mathbb{R}^C \to [0, \infty)$ are the *potential functions*.

Example 2: If the following graph is a graphical model of random variables $X = (X_1, X_2, X_3, X_4)$:



then

$$p_X(x_1, x_2, x_3, x_4) = \psi_{12}(x_1, x_2) \times \psi_{23}(x_2, x_3)\psi_{34}(x_3, x_4)\psi_{14}(x_1, x_4).$$

A *path* in G is a sequence of distinct nodes $v_0, v_1, \ldots, v_n$ s.t. there is an edge between any two consecutive nodes, $v_{i-1} - v_i$ for $i = 1, \ldots, n$. Let $A, B, C \subset V$ be subsets of nodes. Then C *separates* A and B if every path from a node $v \in A$ to a node $w \in B$ intersects C. For instance, in example 1, X_2 separates X_1 and (X_3, X_4) and in example 2, (X_2, X_4) separates X_1 and X_3.

In graphical model, the notion of separation and conditional independence are related via the following theorem.

Theorem 2.3 *Suppose the distribution of $X = (X_v : v \in V)$ factorizes over $G = (V, E)$. Let $A, B, C \subset V$ be subsets of nodes. Then*

$$C \ \textit{separates} \ A \ \textit{and} \ B \implies X_A \perp\!\!\!\perp X_B \mid X_C.$$

The above is a gentle introduction on the graphical model. There will be more about it in the 517, including the famouse Hammersley-Clifford theorem that describes the sufficient and necessary conditions of undirected graphical model.