STAT 491/516: Stochastic Modeling **Autumn 2026**

Lecture 2: Graphical Models

Instructor: Yen-Chi Chen

(♠♠♠) = Contents for graduate students.

2.1 Introduction

Graphical model is an important topic in the modern statistical and machine learning research. The graphical model associate a graph to the probability model of a random vector via conditional independence.

2.2 Conditional Independence

For three RVs X, Y, and Z, we say X, Y are conditional independent given Z (can be a random vector) 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

$$X \perp Y | Z$$

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

Note that $X \perp Y | Z$ 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 2.1 *Let p_{XYZ} be the joint PDF/PMF of X, Y, and Z. Then the followings are equivalent:*

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

(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):

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

■

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

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

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

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

(C4) (contraction)

$$X \perp Y|Z \text{ and } X \perp W|(Y,Z) \iff X \perp (W,Y)|Z.$$

(C5) If the joint PDF $p_{XYZW}(x,y,z,w)$ satisfies $f_{YZW}(y,z,w) > 0$ almost everywhere. Then

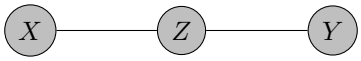
$$X \perp Y|(W,Z) \text{ and } X \perp W|(Y,Z) \iff X \perp (W,Y)|Z.$$

2.3 Undirected graphs

A *graphical model* uses a graph to represent the conditional independence between a set of RVs. We start with the concepts of graphical models and later we will discuss how this model is constructed. Suppose that $X \perp Y|Z$ then we have

$$p_{XYZ}(x, y, z) = p(x, y|z)p(z) = p(x|z)p(y|z)p(z) = g(x, z)h(y, z)$$

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.

The above is the basic definition of a graphical model. We now discuss how this model is constructed. The graphical model relies on two properties: graph factorization (how the distribution of a random variable is associated with a graph) and Markov properties (how the graph represents conditional independence).

2.3.1 Graph factorization and clique decomposition

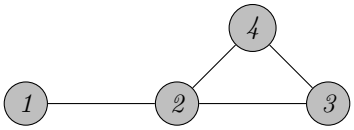
A graph G and a random vector X may or may not have any relationship. The notion of graph factorization connects the joint PDF/PMF of X using a graph G.

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 2.2 *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. In this case, we often call it a maximal clique. We denote the family of all complete sets/maximal cliques as $\mathcal{C}(G)$.

In the above example, complete sets/cliques are

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

And maximal cliques are $\{1, 2\}$, $\{2, 3, 4\}$.

Definition 2.3 *Let $X = (X_1, \cdots, X_d)$ be a random vector and $G = (V, E)$ be a graph where $V = \{V_1, \cdots, V_d\}$ is the node set. We say that X that* **factorizes over/with respect to a graph** *G if there exists (potential) functions $\{\psi_C \geq 0 : C \in \mathcal{C}(G)\}$ such that*

$$p(x_1, \cdots, x_d) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \psi_C(x_C)$$

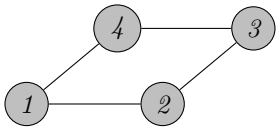
and $Z = \int \prod_{C \in \mathcal{C}(G)} \psi_C(x_C) dx_1, \cdots dx_d$ is known as the partition function.

Note that we call the distribution of X a *Gibbs distribution* with respect to G if

$$p(x_1, \cdots, x_d) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \psi_C(x_C) = \frac{1}{Z} \exp\left(\sum_{C \in \mathcal{C}(G)} \log \psi_C(x_C) \right)$$

for some positive functions $\{\psi_C > 0 : C \in \mathcal{C}(G)\}$.

Example 2.4 *If the following graph is a graphical model of random variables $X = (X_1, X_2, X_3, X_4)$:*



then

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

Definition 2.3 defines the meaning of graph factorization that connects the distribution of a random vector X to a graph G. However, it does not imply anything about the conditional independence. The graph factorization and conditional independence are associated via the Markov properties of graphs.

2.3.2 Global Markov property

The graph factorization leads to conditional independence. Here is an example.

Example 2.5 *We consider the graphical model in Example 2.4 and assume the following setups. All random variables $X_i \in [0, 1]$ and*

$$\begin{aligned}
\psi_{12}(x_1, x_2) &= e^{x_1 x_2} \\
\psi_{23}(x_2, x_3) &= e^{2x_2 x_3} \\
\psi_{34}(x_3, x_4) &= e^{x_3^2 x_4} \\
\psi_{14}(x_1, x_4) &= e^{3x_1 x_4^2}
\end{aligned}$$

The joint density is given by the product of these potential functions normalized by the partition function Z:

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \exp\left(x_1 x_2 + 2x_2 x_3 + x_3^2 x_4 + 3x_1 x_4^2\right)$$

Moreover, it shows some conditional independence. Specifically, we have $X_1 \perp X_3 \mid X_2, X_4$. *To see this, recall that*

$$p(x_1, x_2, x_3, x_4) = \underbrace{\left[\frac{1}{Z}\exp(x_1x_2 + 3x_1x_4^2)\right]}_{g(x_1,x_2,x_4)} \underbrace{\left[\exp(2x_2x_3 + x_3^2x_4)\right]}_{h(x_3,x_2,x_4)}$$

Because the density factorizes into a term g *containing* x_1 *(but not* x_3*) and a term* h *containing* x_3 *(but not* x_1*), we conclude* $X_1 \perp X_3 \mid X_2, X_4$ *by property (v) of Theorem 2.1. However, if we drop one variable in the conditioning, we lose the conditional independence. To check conditional independence given* only X_2*, we must first find the marginal-conditional distribution by integrating out the unobserved variable* X_4.

$$p(x_1, x_2, x_3) = \int_0^1 p(x_1, x_2, x_3, x_4)dx_4$$

$$p(x_1, x_2, x_3) = \frac{1}{Z}\exp(x_1x_2 + 2x_2x_3)\int_0^1 \exp\left(x_3^2x_4 + 3x_1x_4^2\right)dx_4$$

Let $I(x_1, x_3) = \int_0^1 \exp\left(x_3^2x_4 + 3x_1x_4^2\right)dx_4$. *Because* x_1 *and* x_3 *are coupled together through their interaction with the integration variable* x_4*, the resulting function* $I(x_1, x_3)$ *cannot be factored into the form* $A(x_1)B(x_3)$. *Therefore,* $p(x_1, x_2, x_3)$ *cannot be factored into* $g(x_1, x_2)h(x_3, x_2)$. *This implies that* X_1 *and* X_3 *are conditionally dependent given only* X_2. *To render them independent, we must also condition on the path through* X_4.

2.3.2.1 Markov random field and global Markov property

A **Markov random field** is a random variable X satisfying Markov properties (conditional independence) with respect to a graph. It turns out that there are three common Markov properties that associates the graph factorization to the notion of conditional independence. We start with the most common type of Markov properties–*global Markov property.*

The global Markov property relies on the notion of path and separation of a graph. A *path* in G is a sequence of distinct nodes $v_0, v_1, \ldots, v_d$ 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.

Definition 2.6 (Global Markov Property) *A probability distribution* P *for a random vector* $X = (X_1, \cdots, X_d)$ *satisfies the global Markov property with respect to a graph* G *if for any disjoint vertex subsets* A, B*, and* C *such that* C *separates* A *and* B*, then the random variables* X_A *are conditionally independent of* X_B *given* X_C.

It is very easy to see that

Graph Factorization in Definition 2.3 $\Rightarrow$ Global Markov Property

as stated in the following theorem.

Theorem 2.7 (Global Markov theory) *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*

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

A distribution that satisfies the global Markov property is said to be a *Markov random field* or *Markov network* with respect to the graph.

Example 2.8 (Information Spread) *Consider the graph $X - Z - Y$. Suppose these variables represent a rumor or piece of information passed along a chain of people.*

- *X: Person 1 knows the rumor (1 if yes, 0 if no).*
- *Z: Person 2 knows the rumor.*
- *Y: Person 3 knows the rumor.*

If Person 1 only talks to Person 2, and Person 2 only talks to Person 3, the graph is a linear chain. The potential functions, $\psi_{XZ}(x,z)$ and $\psi_{ZY}(z,y)$, represent the reliability of communication between the pairs. Intuitively, if you do not know what Person 2 heard, Person 1's knowledge gives you predictive power about Person 3's knowledge (they are dependent). However, if you know exactly what Person 2 heard (Z is observed), then learning what Person 1 originally said (X) gives you no additional information about what Person 3 will hear (Y). This perfectly illustrates the Markov property: $X \perp Y \mid Z$. The joint PMF factorizes over the maximal cliques $\{X, Z\}$ and $\{Z, Y\}$ as:

$$p(x,z,y) = \frac{1}{Z_{norm}} \psi_{XZ}(x,z)\psi_{ZY}(z,y)$$

2.3.3 Other Markov properties and Hammersley-Clifford theorem (♠♠♠)

Global Markov property is not the only Markov property that we can associate a graph to the probability model. Here we introduce two other Markov properties.

Definition 2.9 (Local Markov Property) *A probability distribution P for a random vector $X = (X_1, \cdots, X_d)$ satisfies the local Markov property with respect to a graph G if the conditional distribution of a variable given all its neighbor is independent of any other vertices. Namely, let $N(j) = \{X_i : E_{ij} = 1\}$ be the neighbors of X_j. Then the local Markov property means that*

$$P(X_j | X_{-j}) = P(X_j | X_{N(j)}),$$

where $X_{-j} = \{X_i : i \neq j\}$.

A more general definition is the pairwise Markov property.

Definition 2.10 (Pairwise Markov Property) *A probability distribution P for a random vector $X = (X_1, \cdots, X_d)$ satisfies the pairwise Markov property with respect to a graph G if for any two non-adjacent vertices X_i and X_j (i.e., $E_ij = 0$),*

$$X_i \perp X_j | X_{V \setminus \{i,j\}}.$$

Proposition 2.11 (Equivalence of Markov properties) *For any undirected graph G and any distribution P, we have*

$$\textit{Global Markov Property} \Rightarrow \textit{Local Markov Property} \Rightarrow \textit{Pairwise Markov Property.}$$

The proof is very straight forward so we omit it.

Example 2.12 (Local Markov property but no global Markov property) *Define binary random variables $X_1, \cdots, X_5$ such that $P(X_1 = 1) = P(X_5 = 1) = \frac{1}{2}$ and $X_2 = X_1$ and $X_4 = X_5$ and $X_3 = X_2 X_4$. You can easily verify that the random vector satisfies the local Markov property with respect to the chain graph G_0 that is $(X_1 - X_2 - X_3 - X_4 - X_5)$. In particular, the PMF of X_3 is conditionally independent of X_1 and X_5 given X_2 and X_4. However, the global Markov property is violated. To see this, consider the case of $X_3 = 0$ and it is easy to see that $P(x_2, x_4 | X_3 = 0) = \frac{1}{3}$ when $(x_2, x_4) = (1, 0), (0, 1), (0, 0)$. However, the marginal probability $P(X_2 = 0 | X_3 = 0) = P(X_4 = 0 | X_3 = 0) = \frac{2}{3}$. Thus,*

$$P(X_2 = 0, X_4 = 0 | X_3 = 0) = \frac{1}{3} \neq P(X_2 = 0 | X_3 = 0) \times P(X_4 = 0 | X_3 = 0) = \frac{4}{9}$$

so the global Markov property does not hold for this graph G_0.

Example 2.13 (Pairwise Markov property but no local Markov property) *Define binary random variables X_1, X_2, X_3 and $X_1 = X_2 = X_3$ with $P(X_1 = 1) = \frac{1}{2}$. The random vector $X = (X_1, X_2, X_3)$ has a very degenerated PMF. Consider a graph G such that there is only one edge $E_{23} = 1$. Then you can easily verify that X satisfies the pairwise Markov property with respect to G but not the local Markov property (specifically, $P(X_1 = 1 | X_2 = 0, X_3 = 0) = 0 \neq P(X_1 = 1) = \frac{1}{2}$). This example also shows a fact about the Markov properties– the same distribution may satisfy* different Markov properties on different graphs*! In the above example, the same pairwise Markov property holds for another graph G' with only a single edge $E'_{12} = 1$ or a graph G'' with only a single edge $E''_{13} = 1$.*

From the above examples, we see that the Markov random fields is not a well-defined term. We have to specifically state which Markov property that a random variable X satisfies with respect to a graph.

A good news is that when the PDF/PMF is positive, the three Markov properties are equivalent.

Proposition 2.14 *Let p be the PDF/PMF of random variable $X \in \mathbb{R}^d$ such that $\mathcal{X}_j$ is the support of X_j for $j = 1, \cdots, d$. If $p(x) > 0$ for all $x \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$, then the three Markov properties are equivalent.*

The condition

$$p(x) > 0 \text{ for all } x \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d \tag{2.1}$$

is also known as the positivity condition (for graphical models).

Proposition 2.14 is a powerful result because it shows that under the positivity condition, all three Markov properties are equivalent. So the term Markov random fields is well-defined.

The above proposition relies on the intersection lemma from

> Pearl, J., & Paz, A. (1985). Graphoids: A graph-based logic for reasoning about relevance relations. University of California (Los Angeles). Computer Science Department.

Lemma 2.15 (Intersection lemma; Pearl, J., & Paz (1985)) *Suppose that for any subsets $A, B, C, D \subset V$ we have*

$$X_A \perp X_B | X_{C \cup D}, \quad X_A \perp X_C | X_{B \cup D} \Rightarrow X_A \perp X_{B \cup C} | X_D.$$

Then the three Markov properties are equivalent.

Proof:[Proof of Proposition 2.14]

Without loss of generality, we consider three variable cases: $X = (X_1, X_2, X_3)$. To use Lemma 2.15, we need to show that

$$X_1 \perp X_2 | X_3, \quad X_1 \perp X_3 | X_2 \Rightarrow X_1 \perp \{X_2, X_3\}.$$

Assume the two conditional independence in the left-hand side of the above equation. Then we have

$$p(x_1, x_2, x_3) = f_{13}(x_1, x_3) f_{23}(x_2, x_3) = g_{12}(x_1, x_2) g_{23}(x_2, x_3)$$

for some functions $f_{13}, f_{23}, g_{12}, g_{23}$. Thus,

$$g_{12}(x_1, x_2) = \frac{f_{13}(x_1, x_3) f_{23}(x_2, x_3)}{g_{23}(x_2, x_3)} = f_{13}(x_1, x_3) \frac{f_{23}(x_2, x_3)}{g_{23}(x_2, x_3)}.$$

An interesting implication from the above equation is that the left-hand side does not depend on x_3 so this holds *for any* x_3. WLOG, we choose $x_3 = 0$ and this leads to

$$g_{12}(x_1, x_2) = f_{13}(x_1, 0) \frac{f_{23}(x_2, 0)}{g_{23}(x_2, 0)} = h(x_1) k(x_2).$$

Putting this back to the joint PDF/PMF, we obtain

$$p(x_1, x_2, x_3) = g_{12}(x_1, x_2) g_{23}(x_2, x_3) = h(x_1) k(x_2) g_{23}(x_2, x_3),$$

which implies $X_1 \perp \{X_2, X_3\}$. So by Lemma 2.15, the three Markov properties are equivalent.

■

Theorem 2.7 shows that if the distribution of a random vector X factorizes over a graph, then it satisfies the global Markov property. However, the reverse direction is unclear to us. Specifically, we want to know

> *if a random vector satisfies the global/local/pairwise Markov property with respect to a graph, can it always be factorized with respect to the graph?*

If we can provide a positive answer to the above question, then we can use graph factorization for a Markov random field. The good news is: the following theorem, known as the Hammersley-Clifford (or Hammersley-Clifford-Besag) theorem, provides a positive answer to this question.

Theorem 2.16 (Hammersley-Clifford (1971)) *Suppose that $G = (V, E)$ is a graph and $X_1, \cdots, X_d$ are random variables that take on a finite number of values. Let $\mathcal{X}_j$ be the support of X_j for $j = 1, \cdots, d$. If $p(x) > 0$ for all $x \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$, and satisfies the local Markov property with respect to G, then it factors with respect to G.*

The Hammersley-Clifford shows that under the positivity condition in equation (2.1), the graph factorization and all three Markov properties are equivalent.

The following paper is the original paper that states this theorem:

> Hammersley, J. M., & Clifford, P. (1971). Markov fields on finite graphs and lattices.

Note that they do not publish this paper in a journal article but you can still find the original manuscript online.

A formal paper that includes this theorem (and improves the proof and mentioned the generalization to continuous random vector) is the following paper:

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 192-225.

Although the Hammersley-Clifford theorem only proves the case for discrete random variables, the result can be generalized to continuous random variables as well. The Hammersley-Clifford theorem together with Proposition 2.14 imply the following conclusion:

For a random vector X with a positive PDF/PMF (satisfying the positivity condition), then

$$\text{satisfying Markov Properties} \Leftrightarrow \text{factorizing with respect to } G.$$

Thus, Theorem 2.7 together with the Hammersley-Clifford theorem provide the foundation of graphical model that we can interchangably use graph factorization and conditional independence. This is why the Hammersley-Clifford theorem is sometimes referred to as *the fundamental theorem of graphical models*.

The Hammersley-Clifford theorem not only informs us that satisfying the three Markov properties is the equivalent to graph factorization, its required condition, the positivity condition in equation (2.1), is the same as the condition for showing equivalence among the three Markov properties. Thus, once a random variable X satisfies the positivity conditions, all three Markov properties are equivalent (Proposition 2.14) and it factorizes with respect to a graph G (Theorem 2.16), so the *Markov random field* is well-defined and factorizes with respect to a graph G.

2.3.4 Gaussian graphical model

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

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

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

In this model, there are two parameters μ and Σ. When Σ is invertible, the PDF satisfies the positivity condition, so X is a well-defined Markov random field with respect to a particular graph G and we can use graph factorization and all Markov definitions.

Now we consider the case where G has no edge between node X_1 and X_2. This implies the conditional independence $X_1 \perp X_2 | X_3, \cdots, X_p$. What does this conditional independence tell us about the underlying parameters?

Using the graph factorization, 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.

Because

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

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.

2.3.5 Log-linear model (♠♠♠)

The log-linear model is a parametrization for the PMF of multinomials. Suppose that each $X_j \in \{0, 1, 2, \cdots, m_j - 1\}$ for each $j = 1, \cdots, d$ and $X = (X_1, \cdots, X_d)$ is the random vector of interest.

- **Log-linear model.** The log-linear model expands the log PMF of X as

$$\log p(x) = \sum_{A \subset V} \psi_A(x_A), \tag{2.2}$$

  with the constraint that if a variable $j \in A$ with $x_j = 0$, $\psi_A(x_A) = 0$. Equation (2.2) is known as the log-linear expansion of $p(x)$. Although $\psi_A(x_A)$ behaves like a function, it is a set of several parameters since the variable(s) x_A only takes discrete values. In fact, there are only $\prod_{j \in A}(m_j - 1)$ number of possible values of ψ_A so it is often referred to as the parameter of a log-linear model. You can interpret the parameter/function ψ_A as the (joint) interaction effect of variables in A. Clearly, when $|\psi_A|$ is finite, the positivity condition holds so there is no ambiguity in the three Markov properties and graph factorization. Thus, the log-linear model is a well-defined Markov random field with respect to a graph G.

- **Hierarchical model.** A *hierarchical log-linear model* is a log-linear model such that if $\psi_A(x_A) = 0$ implies $\psi_B(x_B) = 0$ for all $B \supset A$. Namely, a hierarchical log-linear model has a nested structure that if a parameter $\psi_A = 0$, any parameter that is a superset of A must be 0. You can interpret a hierarchical log-linear model as the model that any higher-order interaction exists only if all lower-order interactions exist. However, even if all lower-order interactions exist, the higher-order interaction needs not to exist.

- **Graphical log-linear model.** A *graphical log-linear model* with respect to a graph G is the log-linear model such that $\psi_A(x_A)$ is not zero if and only if A is a clique (not necessarily a maximal clique). The graphical model requires that if all lower-order interactions exist, the higher-order interaction MUST exist. Thus, one can see that the graphical log-linear model is a sub-class of the hierarchical model, as stated in the following lemma.

Lemma 2.17 *A graphical log-linear model is hierarchical log-linear model but not vice versa.*

Proof:

Suppose that for a graphical model of G with $\psi_A = 0$, this implies that A is not a clique in G. Thus, any set $B \supset A$ will not be a clique in G so the model is hierarchical.

Now consider a three variable log-linear model with

$$\log p(x) = \psi_1(x_1) + \psi_2(x_2) + \psi_3(x_3) + \psi_{12}(x_1, x_2) + \psi_{13}(x_1, x_3) + \psi_{23}(x_2, x_3).$$

Clearly, this is a hierarchical model but not a graphical model (it will requires $\psi_{123}(x_1, x_2, x_3) \neq 0$).

■

With the above lemma, we conclude that

graphical log-linear model ⇒ hierarchical model ⇒ log-linear (multinomial) model

and since log-linear model is a Markov random field when all components are finite, all of the above three models are Markov random field, i.e., they factorize with respect to an undirected graph and satisfying the Markov properties.

Ising model. The Ising model is a special case of hierarchical log-linear models, so it is a Markov random field. It is a hierarchical model with binary variables with only pairwise interactions. Specifically, the Ising model is the case where

$$\log p(x) = \sum_{i=1}^{d} \theta_i x_i + \sum_{(j,k)\in E} \theta_{j,k} x_j x_k. \tag{2.3}$$

Since the Ising model only contains pairwise interaction, it can be viewed as a discrete analogue of the Gaussian graphical model. The Ising model is related to the logistic regression. By the local Markov property, a random variable X_i only depends on its neighborhoods so the conditional probability

$$P(X_i = 1|X_{-i}) = P(X_i = 1|X_j, (i,j) \in E) = \frac{\exp(\theta_i + \sum_{(i,j)\in E} \theta_{i,j} x_j)}{1 + \exp(\theta_i + \sum_{(i,j)\in E} \theta_{i,j} x_j)},$$

where X_{-i} is the collection of all variables except X_i.

Potts model. The Potts model is a generalized Ising model that allows variables to have m distinct outcomes, i.e., $X_i \in \{0, 1, 2, \cdots, m-1\}$ and the pairwise interaction contributes only if the two variables are in the same 'state'. Specifically, the joint PMF in the Potts model can be factorized as

$$\log p(x) = \sum_{i=1}^{d} \theta_i x_i + \sum_{(j,k)\in E} \theta_{j,k} \delta(x_j, x_k), \tag{2.4}$$

where $\delta(a, b) = I(a = b)$. The Potts model is motivated by statistical mechanics in which each variable X_i is a particle and a particle has m different states. In a stable scenario, two adjacent particles (particles are variables X_i's) will avoid being in the same state. So the distribution can be modeled using the Potts model with a negative $\theta_{j,k}$.

2.4 Directed acyclic graphs

A graph where the edges are directional is called a directed graph. In statistics and machine learning, we often focus on one particular directed graph called *directed acyclic graphs (DAGs)*. A DAG is a directed graph that has no directed loops (i.e., arrows do not form a loop). Directed graphical models are often viewed as a generative model. To illustrate the idea, consider 5 random variables $X_1, \cdots, X_5$ with the following generative models:

$$\begin{aligned} X_1 &\sim p_1(x_1) \\ X_2 &\sim p_2(x_2) \\ X_3 &\sim p_3(x_3) \\ X_4|X_1, X_2 &\sim p_4(x_4|X_1, X_2) \\ X_5|X_1, X_3, X_4 &\sim p_5(x_5|X_1, X_3, X_4). \end{aligned}$$

Then we can summarize this model using the left panel of Figure 2.1.

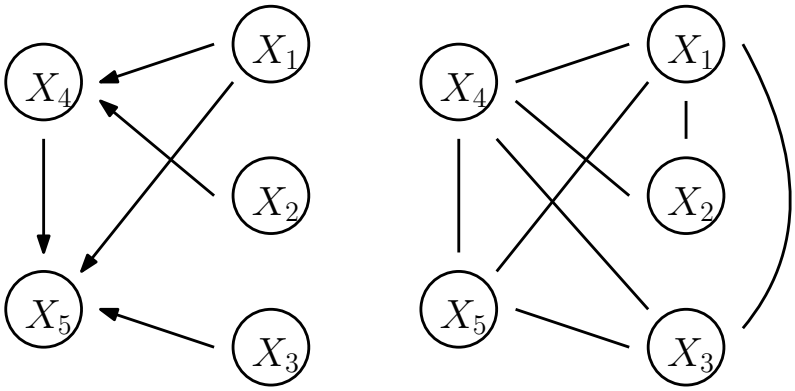


Figure 2.1: **Left:** An example of a DAG with 5 variables. **Right:** The corresponding UG.

Because of the popularity of DAG in the probability generative model, a DAG is also called a Bayesian network. Note that a Bayesian network has nothing to do with Bayesian inference or Bayesian statistics; it is just a graphical model that relied on Bayes rule to describe a probability distribution.

The DAG in the left panel of Figure 2.1 implies that the joint density can be written as

$$p(x_1, \cdots, x_5) = p(x_5|x_1, x_3, x_4)p(x_4|x_1, x_2)p(x_3)p(x_2)p(x_1) = \psi_{1,3,4,5}(x_1, x_3, x_4, x_5)\psi_{1,2,4}(x_1, x_2, x_4)$$

so the corresponding undirected graphical model is the right panel of Figure 2.1 that has two maximal cliques $(1, 3, 4, 5)$ and $(1, 2, 4)$.

More generally, we can always convert a DAG into an UG using the idea of *moralizing*. If there is an arrow from node X_i to node X_j, we call X_i a parent (node) of X_j and X_j a child (node) of X_i. Note that every node may have multiple parents and children.

Definition 2.18 *The moral graph M of a DAG G is an undirected graph where there is an edge between two vertices X_i and X_j if one of the following conditions met:*

- *There is an edge between X_i and X_j in G.*
- *X_i and X_j are the parents of the same child node.*

Informally, the moralized graph can be constructed by 'marrying the parents'–we connect all parents of each child node (and remove arrows) to form the corresponding undirected graph. Two different DAGs may have the same moralized graph, as illustrated in Figure 2.2. Also, the arrow direction matters in the construction of moral graph; Figure 2.3 shows an example that we only reverse one arrow's direction in the DAG of the left panel in Figure 2.2 and the resulting moral graph is different.

Important note on learning DAG. The data alone CANNOT give you a unique DAG! This is because the data can only give you associations (e.g., correlations) among variables. We can have multiple DAGs leading to the same associations. Figure 2.2 is an example of this – we can only identify the dependency as the undirected graph in the right but we are unable to tell which DAG (left or middle) is the actual DAG that generates the data. In actual scientific problem, we have additional knowledge on the variables, which may allow us to choose the correct DAG. But **data alone CANNOT tell you which DAG is the correct one.**

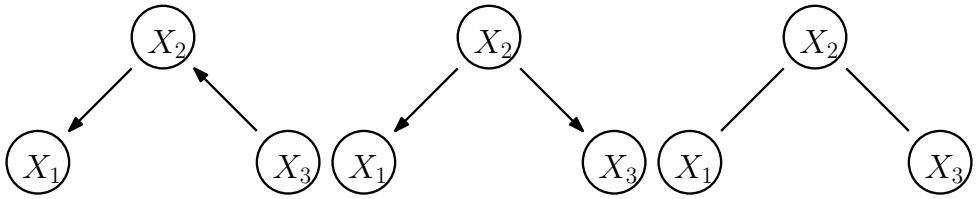


Figure 2.2: **Left and middle:** Two DAGs. **Right:** The moral graph from both DAGs in the left two panels.

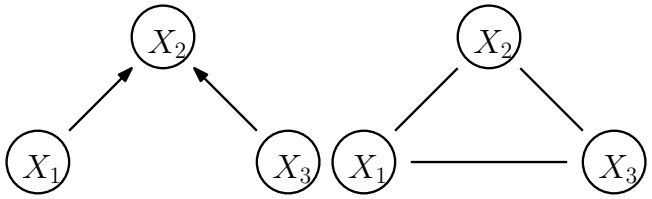


Figure 2.3: **Left:** A DAG that is similar to the left panel as Figure 2.2 but we reverse only one arrow's direction. **Right:** The moral graph from the DAG in the left panel.

2.4.1 d-separation and independence (♠♠♠)

The **d-separation** is common criterion to determine independence or conditional independence in a DAG. Roughly speaking, d-separation implies independence. To explain the concept of d-separation, we first introduce a few concepts.

For a pair of vertices X_a and X_b, an *undirected path* is a collection of vertices $(V_1, \cdots, V_k)$ such that $V_1 = X_a$, and $V_k = X_b$, and each pair (V_j, V_{j+1}) contains an arrow (ignored the orientation). A collider of a path is the triplet (V_{j-1}, V_j, V_{j+1}) of the path such that the arrows are $V_{j-1} \to V_j \leftarrow V_{j+1}$. A vertex X_a is a *descendent* of X_b if there exists a directed path $X_b = V_1 \to V_2 \to \cdots \to V_k = X_a$.

Rule 1: a collider of a path creates a block on the path. If a path has one or more blocks, then the path is d-separated.

For two sets of variables $A, B \subset \{1, 2, \cdots, n\}$, they are d-separated if *all paths between A and B are d-separated.*

If A and B are d-separated, then $X_A \perp X_B$

The Rule 1 is often used to verify the independence between two sets of variables. Now we consider the case of conditional independence, which we will introduce two more rules.

Rule 2: Conditioned on a non-collider will create a block.

Rule 3: Conditioned on a collider or its descendent will remove the block caused by the collider.

For three sets of variables $A, B, C \subset \{1, 2, \cdots, n\}$, A and B are d-separated given C if *all paths between A and B are d-separated* after applying the above rules (Rule 2 and 3 with conditioning on C).

If A and B are d-separated given C, then $X_A \perp X_B | X_C$.

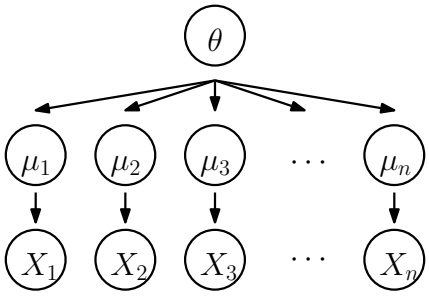


Figure 2.4: A DAG summarizing the relation among random variables $X_1, \cdots, X_n, \mu_1, \cdots, \mu_n, \theta$ described in Section 2.4.2.

Example 2.19 *Consider the graph that*

$$X_1 \to X_2 \to X_3 \leftarrow X_4 \leftarrow X_5.$$

Since X_3 is a collider, we have $X_1 \perp X_5$. However, if we conditioned on X_3, this will remove the block at X_3, making the path between X_1 and X_5 not d-separated. So X_1 and X_5 are not conditional independent given X_3. If we further condition on X_2, this will add an additional block at X_2, making the path d-separated. So we have $X_1 \perp X_5 | X_2, X_3$.

Example 2.20 *Now we consider another graph*

$$Y_1 \to Y_2 \leftarrow Y_3 \to Y_4 \leftarrow Y_5.$$

There are two colliders in the above DAG (Y_2 and Y_4). We now consider the relation between Y_1 and Y_5. Clearly, $Y_1 \perp Y_5$. When we conditioned on Y_2, while this removes the block at Y_2, we still have another collider Y_4. So we still have $Y_1 \perp Y_5 | Y_2$ and similarly $Y_1 \perp Y_5 | Y_4$ Note that Y_1 and Y_5 will be dependent if we conditioned on Y_2 and Y_4. However, since Y_3 is not a collider, conditioned on Y_3 will create a block, making the path d-separated. So we have $Y_1 \perp Y_5 | Y_3$ and even $Y_1 \perp Y_5 | Y_2, Y_3, Y_4$. As long as there is one block on the path, the path will be d-separated.

2.4.2 Hierarchical Bayes (♠♠♠)

A Bayesian hierarchical model is scenario that DAGs are often applied to. To illustrate the idea, we consider the following example. Suppose that we have n individuals participating in an exam and their scores can be summarizes using univariate random variables $X_1, \cdots, X_n$. We all know that scores are measurements (with noises) of the individual's capability so we can view each random variable as

$$X_i | \mu_i \sim N(\mu_i, \sigma^2),$$

where μ_i can be interpreted as the individual's actual performance on the exam. Suppose that these n individuals are randomly chosen from a population. To model the randomness of the selection, we assume that

$$\mu_1, \cdots, \mu_n \sim N(\theta, \tau^2),$$

where θ reflects the average performance of the sampled population. To account for our uncertainty about θ, we may introduce a prior $\pi(\theta)$ over it. Under this model specification, all random quantities can be written as the DAG in Figure 2.4.

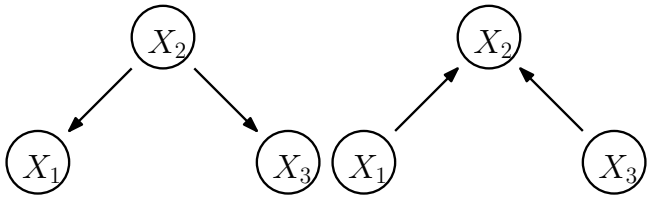


Figure 2.5: Two DAGs.

2.4.3 Causal graph (♠♠♠)

The DAG is also used frequently in causal inference. The arrow is interpreted as a causal relation. For instance, if we have a DAG $X_1 \to X_2 \to X_3$, then we mean that X_1 causes X_2 and X_2 causes X_3. The above graph also implies that conditioning on X_2, X_1 and X_3 are independent. In the causal relation, this means that if we controlled X_2, then X_1 does not causal any change in X_3. So the conditional independence becomes an elegant mathematical tool to discuss causal relation.

Here is another example to illustrate how DAGs provide useful insights on causal relation. Consider the left DAG in Figure 2.5–

- *Causal interpretation:* X_2 causes both X_1 and X_3. Thus, if X_2 is unobserved, then X_1 and X_3 are associated (in this case, X_2 is a *confounder*). On the other hand, if X_2 is controlled, then X_1 and X_3 are independent.
- *Graphical model interpretation:* The generative model is

  $$p(x_1, x_2, x_3) = p(x_1|x_2)p(x_3|x_2)p(x_2).$$

  Thus, the marginal density

  $$p(x_1, x_3) = \int p(x_1, x_2, x_3)dx_2 = \int p(x_1|x_2)p(x_3|x_2)p(x_2)dx_2 = g(x_1, x_3)$$

  for some function g. Thus, X_1 and X_3 are marginally dependent. However, $p(x_1, x_3|x_2) = p(x_1|x_2)p(x_3|x_2)$ so X_1 and X_3 are conditionally independent.

Now we consider the right DAG in Figure 2.5–

- *Causal interpretation:* Both X_1 and X_3 causes X_2 but they are independent causes. However, if X_2 is observed, then X_1 and X_3 will be associated. Note that X_2 in this case will be called a *collider*.
- *Graphical model interpretation:* The generative model is

  $$p(x_1, x_2, x_3) = p(x_2|x_1, x_3)p(x_1)p(x_3) \Rightarrow p(x_1, x_3) = p(x_1)p(x_3)$$

  so X_1 and X_3 are marginally independent. And the conditional density

  $$p(x_1, x_3|x_2) = \frac{p(x_1, x_2, x_3)}{p(x_2)} = \frac{p(x_2|x_1, x_3)p(x_1)p(x_3)}{p(x_2)}$$

  cannot be factorized into the product of $g_1(x_1, x_2)$ and $g_2(x_2, x_3)$ so X_1 and X_3 are conditionally dependent given X_2.

Therefore, the probabilistic structure implied by a DAG and the causal interpretation of variables have an elegant correspondence. This is why DAGs are very popular in causal inference.

2.4.4 Structural Equation Model

The structural equation model (SEM) is a popular model that associate a DAG to a probability distribution via equations. For each variable X_i, we let $\mathsf{PA}(i) \subset V$ denotes the parents of X_i. The structural equation model is a set of generative equations that for each i,

$$X_i = f_i(X_{\mathsf{PA}(i)}, \epsilon_i), \tag{2.5}$$

where ϵ_i are IID random variables with mean 0 and f_i are functions belong certain classes. Equation (2.5) is sometimes called nonparametric SEM, as it does not specify any parametric form.

Linear SEM. Equation (2.5) may run into identification problem if the class of functions are not specified properly. To resolve this issue, people often assume a specific parametric class of each equation. The most popular class is the linear function class, which leads to the famous linear SEM. The linear SEM formulates

$$X_i = \gamma_{i,\mathsf{PA}(i)}^T X_{\mathsf{PA(i)}} + \epsilon_i.$$

Very often people would further assumes that ϵ_i are IID $N(0, \sigma^2)$[1]. Under the linear SEM, you can show that the whole random vector $X \in \mathbb{R}^n$ can be written as

$$X = (\mathbf{I}_n - \Gamma)^{-1}\epsilon,$$

where $\epsilon^T = (\epsilon_1, \cdots, \epsilon_n)$ and Γ is constructed from $\gamma_{i,\mathsf{PA}(i)}$ with $\mathsf{diag}(\Gamma) = (0, 0, \cdots, 0)$ and the feature that if there is no arrow from X_j to X_i, then $\Gamma_{ij} = 0$. Under linear SEM, if $\epsilon \sim N(0, \sigma^2\mathbf{I}_n)$, we further have

$$X \sim N(\mu, \sigma^2(\mathbf{I}_n - \Gamma)^{-1}[(\mathbf{I}_n - \Gamma)^{-1}]^T).$$

The parameters σ^2 and Γ can be estimated by either the MLE or eigen-analysis on the covariance matrix of X.

Example 2.21 (The DAG in Figure 2.1) *We now use the DAG in Figure 2.1 as an example of the linear SEM with Gaussian noises. This DAG implies the following five equations:*

$$\begin{aligned}
X_1 &= \epsilon_1 \\
X_2 &= \epsilon_2 \\
X_3 &= \epsilon_3 \\
X_4 &= \gamma_{4,1}X_1 + \gamma_{4,2}X_2 + \epsilon_4 \\
X_5 &= \gamma_{5,1}X_1 + \gamma_{5,3}X_3 + \gamma_{5,4}X_4 + \epsilon_5.
\end{aligned}$$

This can be written as

$$X = \underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \gamma_{4,1} & \gamma_{4,2} & 0 & 0 & 0 \\ \gamma_{5,1} & 0 & \gamma_{5,3} & \gamma_{5,4} & 0 \end{pmatrix}}_{=\Gamma} X + \epsilon,$$

which after rearrangements, becomes

$$(\mathbf{I}_5 - \Gamma)X = \epsilon \Rightarrow X = (\mathbf{I}_5 - \Gamma)^{-1}\epsilon.$$

[1] You may add the intercept term into the linear SEM. Here we ignore it to obtain an elegant result.

2.4.5 Factor Model/Analysis

Factor analysis is a special type of linear SEM that it includes *latent/hidden/unobserved* variables called factors. Now we consider the simplest factor model that we have n observed variables $X = (X_1, \cdots, X_n)^T$ and k latent factors $G = (G_1, \cdots, G_k)^T$. In this simplest case, the factor model assumes a graphical model with arrows from every G_j to every X_i, which under the linear SEM, leads to the following equation

$$X_i = \sum_{j=1}^{k} \gamma_{i,j} G_j + \epsilon_i \tag{2.6}$$

for each i and G_j, ϵ_i are independent for every pair (i, j). We can write equation (2.6) as

$$X = \Gamma G + \epsilon$$

with $\mathbb{E}(\epsilon) = 0, \mathsf{Cov}(\epsilon) = \sigma^2 \mathbf{I}_n$. Since the latent variables G are unobserved, we often assume that $\mathbb{E}(G) = 0, \mathsf{Cov}(G) = \mathbf{I}_n$ for identifiability. Under this assumption, we have

$$\mathsf{Cov}(X) = \Gamma\Gamma^T + \sigma^2 \mathbf{I}_n.$$

Thus, the problem reduces to an eigenanalysis on the covariance matrix, so the principal component analysis (PCA) is a popular method to estimate Γ. Alternatively, we may assume that ϵ, G are both multivariate normal, and then use the MLE (maximum likelihood estimator) to estimate Γ.

There are many variants of the above factor models. For instances, you may use two sets of latent factors (known as the bi-factor model), or pre-specify some 0's in Γ so that not all factors are influencing every observables. You may even include some arrow among latent factors. The factor analysis is very popular in social sciences such as economics, psychology, and education because factors are often from prior scientific knowledge and we can use our knowledge to add/remove arrows among these variables (latent or observed).