STAT 535: Statistical Machine Learning **Autumn 2024**

Lecture 5: Graphs and Networks

Instructor: Yen-Chi Chen

5.1 Introduction

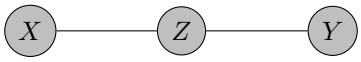
Graphical models and networks models are important topics in the modern statistical and machine learning research. Both methods use the graphs a lot but they are quiet different ideas. In graphical models, graphs are not our data but are tools to determine relationship among entries of a random vector. In network models, the graph are the data (i.e., we observe the network) and we want to make inference with this type of data (known as network data).

5.2 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 $X \perp\!\!\!\perp Y|Z$ 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).

5.2.1 Graph factorization and clique decomposition

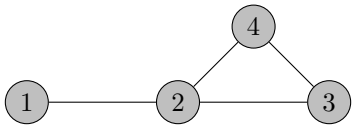
A graph G and a random vector X may or may not have any relationship. The notion of graph factorization connects the joint PDF/PMF of X using a graph G.

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

Definition 5.1 *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*

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

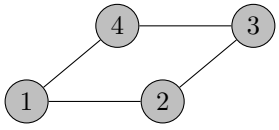
and $Z = \int \prod_{C \in \mathcal{C}(G)} \psi_C(x_C) dx_1, \cdots dx_d$ is known as the partition function.

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: 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 5.1 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.

5.2.2 Markov properties

There are three 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 5.2 (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 the graph factorization in definition (5.1) implies the global Markov property as stated in the following theorem.

Theorem 5.3 (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*

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

A distribution that satisfies the global Markov property is said to be a *Markov random field* or *Markov network* with respect to the graph. A more general type of Markov property is the local Markov property, which is defined as follows.

Definition 5.4 (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 5.5 (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., $E_{ij} = 0$),*

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

Proposition 5.6 (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: 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. 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.

Example: 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 a Markov property 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 $E''_{13} = 1$.

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

Proposition 5.7 *For a distribution P with a PDF/PMF p that is positive, then the three Markov properties are equivalent.*

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

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

Then the three Markov properties are equivalent.

Proof:[Proof of Proposition 5.7]

Without loss of generality, we consider three variable cases: $X = (X_1, X_2, X_3)$. To use Lemma 5.8, we need to show that

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

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 $X_1 \perp\!\!\!\perp \{X_2, X_3\}$. So by Lemma 5.8, the three Markov properties are equivalent.

■

5.2.3 Hammersley-Clifford theorem

Theorem 5.3 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 Markov property with respect to a graph, can it always be factorized with respect to the graph?*

The following theorem, known as the Hammersley-Clifford (or Hammersley-Clifford-Besag) theorem, provides a positive answer to this question.

Theorem 5.9 (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. If $P(x) > 0$ is strictly positive and satisfies the local Markov property with respect to G, then it factors with respect to G.*

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 an be generalized to continuous random variables as well. The Hammersley-Clifford theorem together with Proposition 5.7 imply the following conclusion:

> For a random vector X with a positive PDF/PMF, then
>
> satisfying Markov Properties $\Leftrightarrow$ factorizing with respect to G.

Thus, Theorem 5.3 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*.

5.2.4 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. What does the conditional independence $X_1 \perp\!\!\!\perp X_2 | X_3, \cdots, X_p$ 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.

5.2.5 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. Recall that $G = (V, E)$ is the graph such that $V = \{1, \cdots, d\}$. The log-linear model expands the log PMF of X as

$$\log p(x) = \sum_{A \subset V} \psi_A(x_A), \tag{5.1}$$

with the constraint that if a variable $j \in A$ with $x_j = 0$, $\psi_A(x_A) = 0$. Equation (5.1) 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.

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. Moreover, 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.

Lemma 5.10 *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 model ⇒ hierarchical model ⇒ log-linear (multinomial) model.

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

$$\log p(x) = \sum_{i=1}^{d} \theta_i x_i + \sum_{(j,k)\in E} \theta_{j,k} x_j x_k. \tag{5.2}$$

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

$$\log p(x) = \sum_{i=1}^{d} \theta_i x_i + \sum_{(j,k)\in E} \theta_{j,k} \delta(x_j, x_k), \tag{5.3}$$

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 $\theta_{j,k}$.

5.3 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

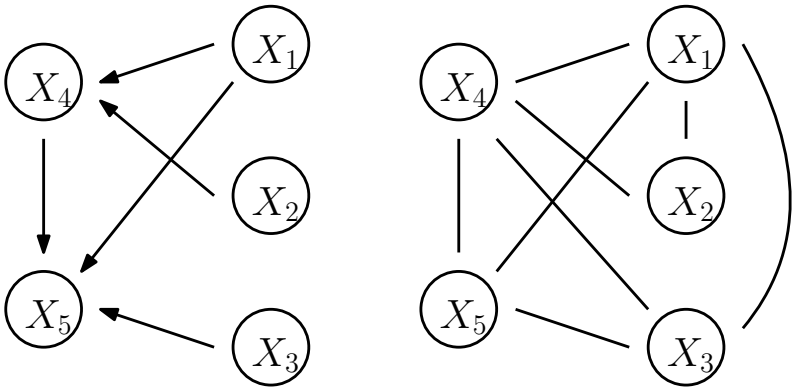


Figure 5.1: **Left:** An example of a DAG with 5 variables. **Right:** The corresponding UG.

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 5.1.

Formally, a random vector X factorized with respect to a DAG G if the joint density

$$p(x_1, \cdots, x_d) = \prod_{j=1}^{d} p(x_j|\mathsf{PA}_{x_j}),$$

where $\mathsf{PA}_{x_j} = \{x_k : \text{ there is an directed arrow/edge from } X_k \text{ to } X_j \text{ in } G\}$ is called the *parent nodes* of X_j in the DAG.

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 5.1 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 5.1 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 note 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 5.11 *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:*

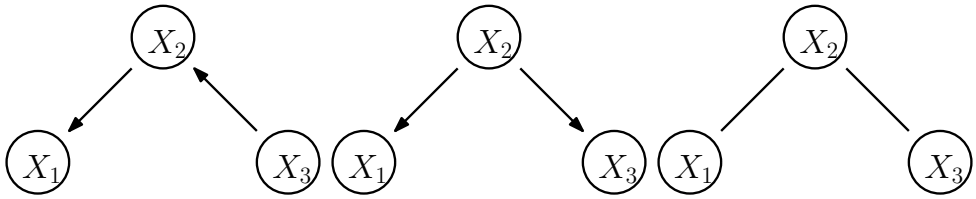


Figure 5.2: **Left and middle:** Two DAGs. **Right:** The moral graph from both DAGs in the left two panels.

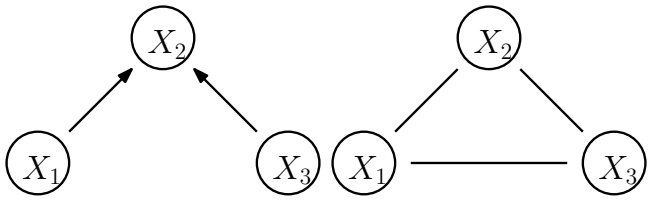


Figure 5.3: **Left:** A DAG that is similar to the left panel as Figure 5.2 but we reverse only one arrow's direction. **Right:** The moral graph from the DAG in the left panel.

- *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 5.2. Also, the arrow direction matters in the construction of moral graph; Figure 5.3 shows an example that we only reverse one arrow's direction in the DAG of the left panel in Figure 5.2 and the resulting moral graph is different.

5.3.1 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 univaraite random variables $X_1, \cdots, X_n$. We all know that scores are measurements

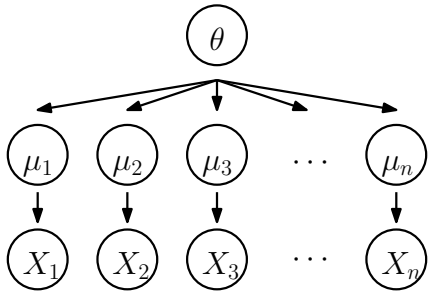


Figure 5.4: A DAG summarizing the relation among random variables $X_1, \cdots, X_n, \mu_1, \cdots, \mu_n, \theta$ described in Section 5.3.1.

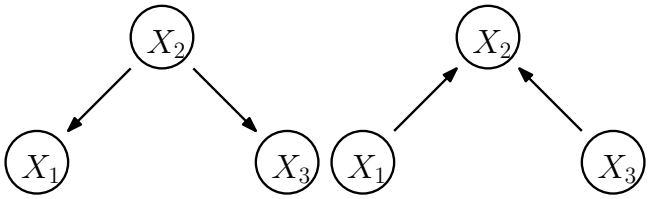


Figure 5.5: Two DAGs.

(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 5.4.

5.3.2 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 5.5–

- *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 5.5–

- *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 *collidor*.

- *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.

5.4 Statistical network models

I would recommend the following lecture notes if you are interested in learning more about network models in Statistics:

> CMU 36-720, Statistical Network Models (by C. Shalizi): https://www.stat.cmu.edu/~cshalizi/networks/16-1/
> UW CSSS-STAT 567 Statistical Analysis of Social Networks (by P. Hoff): https://www.stat.washington.edu/people/pdhoff/courses/567/

In Statistics, networks models are often used to model a network data. Unlike the graphical model problems, in handling the network data, we directly observe a network. Studies on statistical network models attempt to use network data to make scientific inference. There are several scenarios that a network data can be used in statistical inference, for instance

- **Random networks.** We may view the network as random quantities (called random networks) and study the distribution that generate a random network.
- **Community detection.** We want to find communities (nodes are highly interconnected) within a network–these communities are often represent certain groups of nodes.
- **Networks as covariates.** In some scenarios, we may use the network as a covariate in a regression/classification task.
- **Sampling a network data.** In many realistic situation, we may not observe the complete network data but only a fraction of it. Different sampling scheme in this case leads to a different estimator of the properties of the entire network.

5.4.1 Random networks

A statistical network model is a probability model that describes the generating process of a *random graph*. Often the model describes the probability structure of a random undirected and unweighted graph although many model can be generalized to directed graphs as well. In a network model, the nodes are often assumed to be fixed and non-random and the edges are randomly formed (although this is not strict–there are network

models that nodes can be randomly generated). One of the most famous network model is the Erdos-Renyi graph, which states that any pair of note has equal probability to form an edge. A generalized version of the Erdos-Renyi graph is the *stochastic block model* where we assume that all nodes can be partitioned into K unknown groups and pairs of nodes has different probability forming an edge depending on if they belong to the same group or not. We will briefly review and discuss some famous network models. For undirected and unweighted random networks, the probability model is equivalent to an $n \times n$ random matrix with Bernoulli random variable in every entry. So the random network is a special case of a random matrix. Let $G = (V, E)$ be a random graph and $\|V\| = n$ is the number of vertices and we may use the edge/adjacency matrix $E \in \{0,1\}^{n \times n}$ to denote the edges with $E_{ij} = 1$ means that there exists an edge between node i and node j.

Erdös-Rényi model. The Erdös-Rényi model is a very simple stochastic model for generating a random graph. There are two variants of the Erdös-Rényi model.

- $ER(n, p)$ model. This variant is the model that the Erdös-Rényi model is the most commonly referred to as. It generates a random graph that every possible edge has an independent probability of p to form. Namely, $P(E_{ij} = 1) = p$ and $\{E_{ij} : i \geq j\}$ are IID. Essentially, its randomness can be described by $\binom{n}{2}$ independent Bernoulli random variables. This models has several interesting properties on the asymptotic behavior of n and p, for instance,

  1. If $np < 1$, then the graph will almost surely has no connected components with a size larger than $O(\log n)$.
  2. If $np = 1$, then the graph will almost surely has the largest connected component with a size at the order of $O(n^{2/3})$.
  3. If $np \to c > 1$, then the graph will almost surely has a giant connected component and no other connected component has size larger than $O(\log n)$.
  4. If $p < \frac{(1-\epsilon)\log n}{n}$ for some fixed number $\epsilon > 0$, then the graph will almost surely be disconnected.
  5. If $p > \frac{(1+\epsilon)\log n}{n}$ for some fixed number $\epsilon > 0$, then the graph will almost surely be (fullly) connected.

  The above results are summarized from the following famous paper by Erdös and Rényi:

  > Erdös, Paul, and Alfréd Rényi . "On the evolution of random graphs." *Publ. Math. Inst. Hung. Acad. Sci* 5, no. 1 (1960): 17-60.

- $ER(n, m)$ model. This variant creates a random graph with a fixed amount of edges. Here, n stands for the number of vertices and m stands for the total number of edges. The $ER(n, m)$ model generates a random graph such that any graph with m edges has an equal probability being selected.

In most cases, the Erdös-Rényi model refers to the first variant. A simple statistical estimator of p is

$$\widehat{p} = \frac{1}{\binom{n}{2}} \sum_{i \neq j} E_{ij},$$

the fraction of existing edges. However, there is a very sad news about this model–most of the observed networks are not from Erdös-Rényi model. The first three properties partition the possible range of n, p into three categories and from this model, it is unlikely that the graph will have multiple 'stars' (stars refer to the vertices that have high degrees, i.e., many other nodes connecting to them). Many realistic networks such as the social networks often contain several stars.

Stochastic block model (SBM). A popular alternative to the Erdös-Rényi model is the stochastic block model. The idea is very simple. Suppose that there is a partition of the vertices that forms K groups of

vertices. The SBM places an equal probability for forming a within-group edge and another equal probability for forming a between-group edge. And every edge is formed independently from each other. This probability model has $K(K+1)/2$ parameters and the parameters form a symmetric $K \times K$ matrix θ such that the diagonal describes the probability of forming a within-group edge in each group and the off-diagonal parts are the probability of forming a between-group edge that corresponds to the two groups. Let $g_i \in \{1, 2, \cdots, K\}$ denotes the group that the node i belongs to. Then the stochastic block model can be written as

$$P(E_{ij} = 1) = \theta_{g_i, g_j}.$$

If we rearrange the vertices such that vertices are ordered with respect to the index of the nodes, you can easily see that the probability matrix $P(E_{ij} = 1)$ forms a block-diagonal structure. Often the label of which group that a vertex belongs to is unknown and the parameters are also unknown. To estimate the parameter, we can estimate the parameters by the MLE but the likelihood function is often non-convex so finding the MLE is computationally challenging. Often people use approximation approach to find a surrogate of the MLE. One common approximation is the the variational approximation (also known as the variational inference/variational Bayes). See the following papers for more details:

1. Bickel, Peter, et al. "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels." *The Annals of Statistics* 41.4 (2013): 1922-1943.
2. Celisse, Alain, Jean-Jacques Daudin, and Laurent Pierre. "Consistency of maximum-likelihood and variational estimators in the stochastic block model." *Electronic Journal of Statistics* 6 (2012): 1847-1899.

Another approach to estimate the parameter is via the spectral clustering algorithm. The idea is due to the fact that spectral clustering is a relaxation of the optimal graph cut problem and the optimal graph cut is a good approximation to the partition that forms the group structures in SBM. Thus, the spectral clustering gives a partition of the graph and after forming the partition, we can simply use the average within/between-group edge proportion as an estimator of the parameter. See the following for more details:

Rohe, Karl, Sourav Chatterjee, and Bin Yu. "Spectral clustering and the high-dimensional stochastic blockmodel." *The Annals of Statistics* 39, no. 4 (2011): 1878-1915.

Random dot product graph (RDPG). Random dot product graph assumes

$$P(E_{ij} = 1) = X_i^T X_j$$

with X_i, X_j are supported on the ball $\mathbb{S}_d = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and $X_1, \cdots, X_n$ are assumed to be IID from an unknown distribution F over $\mathbb{S}_d$ and denote $X \in \mathbb{R}^{n \times d}$ be the matrix of $X_1, \cdots, X_n$. Interestingly, under RDPG, there exists a simple approach to recover the latent position up to rotations called the *Adjacency spectral embedding (ASE)*. Given an observed edge matrix E, let $\Omega_d \in \mathbb{R}^{d \times d}$ be the diagonal matrix consists of the top d eigenvalues and $U_d \in \mathbb{R}^{n \times d}$ be the corresponding eigenvector matrix. Define a matrix $\widehat{X} = U_d \Omega_d^{1/2} \in \mathbb{R}^{n \times d}$. Then $\widehat{X}$ is a consistent estimator of the latent position matrix X in the sense that there exists an orthogonal matrix $Q \in \mathbb{R}^{n \times n}$ such that

$$\max_i \|Q\widehat{X}_i - X_i\| \leq \frac{C \log n}{n}$$

with a probability of at least $1 - Cn^{-2}$. This result is from the following paper:

Lyzinski, Vince, et al. "Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding." *Electronic journal of statistics* 8.2 (2014): 2905-2922.

For a recent review/survey on RDPG, please see

> Athreya, Avanti, et al. "Statistical inference on random dot product graphs: a survey." *The Journal of Machine Learning Research* 18.1 (2017): 8393-8484.

Latent space model. Both SBM and RDPG are latent space models. The latent space model assumes that there is a latent space $\mathbb{S} \subset \mathbb{R}^d$ such that every node V_i has a latent position $X_i \in \mathbb{S}$. And the probability of forming an edge between node i and j depends on their relative position in the latent space. Namely,

$$P(E_{ij} = 1) = \mu(X_i, X_j).$$

A simple choice is $\mu(x, y) = \log\text{odds}(\alpha + \|x - y\|)$, namely,

$$P(E_{ij} = 1) = \log\text{odds}(\alpha + \|X_i - X_j\|) = \log\text{odds}(\alpha + A_{ij}),$$

where α is a parameter. The estimation of the parameter α and the matrix A_{ij} are often done by the ML procedure but this could be computationally challenging. Note that we can only recover A_{ij} rather than the exact location X_i and X_j because the model will be translational and rotational invariant with respect to X_i's. For latent space model for networks, I would recommend the first paper on this topic:

> Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. "Latent space approaches to social network analysis." *Journal of the american Statistical association* 97.460 (2002): 1090-1098.

Exponential family Random Graph Model (ERGM). The ERGM utilizes the exponential family in statistics to model the generating probability of a specific graph. It does not assume any independence between pairs of edges so it is a very flexible and powerful model. Recall that to describe the probability model of the network, we only need to specify the randomness of edges or the edge matrix E. The ERGM models the probability of the random matrix E as

$$P(E = e; \theta) \propto \exp\left(\sum_{\ell=1}^{d} \theta_\ell T_\ell(e)\right) = \exp\left(\theta^T T(e)\right),$$

where $e \in \{0,1\}^{n\times n}$ is a realization of the edge matrix and $T(e) = (T_1(e), \cdots, T_d(e))$ are the sufficient statistics of the model and $\theta = (\theta_1, \cdots, \theta_d)$ is the parameter. Sometimes, we will introduce the partition function $Z(\theta)$ and write the above model as

$$P(E = e; \theta) = \frac{1}{Z(\theta)} \exp\left(\theta^T T(e)\right),$$

where

$$Z(\theta) = \sum_e \exp\left(\theta^T T(e)\right).$$

The Erdös-Rényi model is a special case of ERGM. To see this, note that from Erdös-Rényi model,

$$\begin{aligned} P(E = e) = \prod_{i,j=1}^{n} p^{e_{ij}}(1-p)^{1-E_{ij}} &= \exp\left(\sum_{i,j} e_{ij}\log p + \sum_{i,j}(1 - e_{ij})\log(1-p)\right) \\ &= \exp\left(n^2\log(1-p) + \sum_{i,j} e_{ij}\log\left(\frac{p}{1-p}\right)\right). \end{aligned}$$

So a sufficient statistic is $T(e) = \sum_{i,j} e_{ij}$, the total number of edges.

In ERGM, all we need is to estimate the parameter θ. In a usual exponential family, this is often done by the MLE. Given a random adjacency matrix E, you can actually show that the MLE of θ_ℓ satisfies

$$T_\ell(E) = \mathbb{E}_{E' \sim P_{\hat{\theta}}}(T_\ell(E')).$$

The left-hand-side is a fixed sufficient statistic and all we need to do is to find the parameter such that the expected value of the sufficient statistic happens to be the same as the observed sufficient statistic. Although this seems to be not difficult (in the regular exponential family problem), it is actually a computationally challenging problem. For any parameter θ, we do not have a closed-form of $\mathbb{E}_{E' \sim P_\theta}(T_\ell(E'))$ so the only way to compute the expectation is to compute the probability $P(E'; \theta)$ for each possible E' and then use the fact

$$\mathbb{E}_{E' \sim P_\theta}(T_\ell(E')) = \sum_e T_\ell(e) P(E = e; \theta).$$

The problem of the above summation is that there are too many terms in the summation of e. For a network with n vertices, there are $2^{\binom{n}{2}}$ possible edges. So it is almost impossible to compute the sum when n is not small. There are some stochastic approximations or MCMC methods for approximating the MLE. For instance, the `statnet`[1] is an R package developed at our department to address this issue.

Statistical inference. Recently, there is more and more attentions on making statistical inference with network models. But in general, it is hard to construct a confidence interval for a parameter of interest under a network model. And it is not so easy to construct a resampling method such as the bootstrap in the network models. However, there are some progress on the resampling inference from a network model. One possible approach is a parametric bootstrap: we assume a parametric model (such as SBMs) and resample from the fitted parametric model; see, e.g.,

> Bickel, Peter, et al. "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels." *The Annals of Statistics* 41.4 (2013): 1922-1943.

Another approach is to use the latent space model. We resample the fitted latent space positions and regenerate a new network from the resampled latent space positions. See the following paper for an example:

> Levin, Keith, and Elizaveta Levina. "Bootstrapping Networks with Latent Space Structure." arXiv preprint arXiv:1907.10821 (2019).

Incorporating covariates. It is possible to incorporate the covariates into the network model. The ERGMs can easily have some effects from the covariates. For instance, if we have a network representing by an adjacency matrix E and covariates $X \in \mathbb{R}^{n \times p}$ for each node, we may use the logistic regression model

$$P(E = e | X; \beta) = \prod_{i \neq j} \left(\frac{\exp(E_{ij}(\beta_0 + \beta_1^T (X_i + X_j)))}{1 + \exp(\beta_0 + \beta_1^T (X_i + X_j))} \right).$$

See the following lecture note for more details:

> https://www.stat.washington.edu/people/pdhoff/courses/567/Notes/l13_ergmcov.pdf

[1] https://statnet.org/trac/wiki/Resources

5.4.2 Other statistical analysis with networks

Community detection. The goal of community detection is to find communities within a network data. While there are several definitions of communities, often a community is a collection of nodes such that nodes share many inter-community connections. In a sense, a community is like a cluster of observations. The stochastic block model is a common model for modeling communities– nodes within the same block are in the same community. A recent survey on this topic can be found in

> Abbe, Emmanuel. "Community detection and stochastic block models: recent developments." *The Journal of Machine Learning Research* 18, no. 1 (2017): 6446-6531.

Networks as covariates. The network may be used as a covariate that informs us the dependency among the response variable. A good news of this use of network is that we no longer have to worry about the probability model that generates the observed networks. For some work along this direction, please see:

> 1. Forastiere, Laura, Edoardo M. Airoldi, and Fabrizia Mealli. "Identification and estimation of treatment and interference effects in observational studies on networks." arXiv preprint arXiv:1609.06245 (2016).
> 2. Basse, Guillaume W., and Edoardo M. Airoldi. "Model-assisted design of experiments in the presence of network-correlated outcomes." *Biometrika* 105, no. 4 (2018): 849-858.
> 3. Basse, Guillaume W., and Edoardo M. Airoldi. "Limitations of design-based causal inference and A/B testing under arbitrary and network interference." *Sociological Methodology* 48, no. 1 (2018): 136-151.

Sampling a network data In many realistic scenarios, we do not observe the entire network but only a fraction of it. For instance, medical researchers often use 'coupons' for participants to recruit other participants. This generates samples from a network structure (assuming that a participant only gives the coupon to his/her friends).

Note that sometimes we can design how the network is sampled but sometimes we cannot design the sampling scheme–we already observed the network. However, even we cannot design the sampling scheme, if we have information about how the network is sampled, we can construct a corresponding probabilistic model that helps us understand properties of an estimator.

There are four common types of sampling a network data:

- *Node-induced subgraph sampling.* We random choose m out of n nodes from the graph and examine if there are edges within these m sampled nodes. This generates a subgraph $G' \subset G$. This is the common scenario that we recruit several participants to join a study and examine their relationships.
- *Edge-induced subgraph sampling.* Randomly choose a set of edges from G and construct the corresponding subgraph. Note that a node will be included in the sampled subgraph if any of its edge is sampled. Again, this generates a subgraph $G' \subset G$.
- *Egocentric sampling.* Similar to the node-induced approach but whenever we observe a node, we also observe all its edges along with adjacent nodes. In addition, if any of the adjacent nodes are linked, we also observe this information. This occurs in questionnaires such as the ones asking
  1. Who are your friends?
  2. Among your friends, who are friends with each other?

- *Link-tracing sampling (also known as snowball sampling).* Link-tracing sampling is similar to the egocentric sampling but we repeat this process several times. We perform egocentric sampling on those who are already recruited and expand the sampled network gradually.

For more details, I would recommend the following lecture note

https://www.stat.washington.edu/people/pdhoff/courses/567/Notes/l18_sampling.pdf

A Properties of Conditional Independence

A.1 Independence Revisited

Recall that two random variables X and Y are independent if

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

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 5.12 *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 $q_{ij} \ge 0$, we may choose every elements of u and v to be non-negative, i.e., $u_j \ge 0$ and $v_j \ge 0$ 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 $v_+ u_+ = \sum_{j=1}^m v_j \sum_{i=1}^n u_i = \sum_{i,j} u_i v_j = \sum_{i,j} p_{ij} = 1$.

■

A.2 Conditional Independence

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

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

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

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

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

Theorem 5.13 *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) ⇒ (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) $X \perp\!\!\!\perp Y|Z \iff Y \perp\!\!\!\perp X|Z$.

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

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

(C4) (contraction)

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

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

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