2023-2024 Graphical Model Notes

Chapter 3 Exponential Families and Contingency Tables

- Denote $X_V \equiv (X_v : v \in V)$, indexed by $V = \{1, \cdots, p\}$. Each X_v takes values in the set $\mathscr{X}_v$. For a subset $A \subseteq V$, we write X_A to denote $(X_v : v \in A)$.

- Let $p(\cdot;\theta)$ be a collection of probability densities over $\mathscr{X}$ indexed by $\theta \in \Theta$. We say that p is an exponential family if it can be written as

$$\begin{aligned} p(x;\theta) &= \exp\left\{\sum_i \theta_i \phi_i(x) - A(\theta) - C(x)\right\} \\ &= \exp\{<\theta, \phi(x)> - A(\theta) - C(x)\} \end{aligned}$$

  - The family is said to be regular if Θ is a non-empty open set.
  - The functions ϕ_i are the sufficient statistics.
  - The components θ_i are the canonical/natural parameters.
  - The function $A(\theta)$ is the cumulant function such that the distribution normalises:

$$A(\theta) = \log \int \exp\{<\theta, \phi(x)> - C(x)\}\,\mathrm{d}x$$

  - The function $Z(\theta) \equiv e^{A(\theta)}$ is the partition function.

- We have

$$\nabla A(\theta) = \mathbb{E}_\theta \phi(X) \qquad \nabla\nabla^{\mathrm{T}} A(\theta) = \mathrm{Cov}_\theta \phi(X)$$

  $A(\theta)$ and $-\log p(x;\theta)$ are convex in θ, and the map $\mu(\theta) : \theta \mapsto \nabla A(\theta)$ is bijective, named the mean function. We care about convexity because it does not have multiple local minima, which in turn facilitates computation.

- Let $X_V = (X_1, \cdots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ be a random vector. Let $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ be a positive definite symmetric matrix. We say that X_V has a multivariate Gaussian distribution with μ and Σ if the joint density is

$$\begin{aligned} f(x_V) &= \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x_V - \mu)^{\mathrm{T}}\Sigma^{-1}(x_V - \mu)\right\} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2}x_V^{\mathrm{T}} K x_V + \mu^{\mathrm{T}} K x_V - \frac{1}{2}\mu^{\mathrm{T}} K \mu + \frac{1}{2}\log|K|\right\} \qquad x_V \in \mathbb{R}^p \end{aligned}$$

  Here, $K \equiv \Sigma^{-1}$ is called the concentration matrix. Let

$$\phi(x_V) = \left(x_v, -\frac{1}{2}x_V x_V^{\mathrm{T}}\right) \qquad \theta = (K\mu, K)$$

  we could easily tell that the multivariate Gaussian distribution is an exponential family[1].

[1] For two matrices A and B, we have $<A, B> = \mathrm{tr}(A, B^{\mathrm{T}})$.

- Let X_V have a multivariate Gaussian distribution with concentration matrix $K = \Sigma^{-1}$, then $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}$ iff $k_{ij} = 0$.
- Let $X_V^{(i)} = (X_1^{(1)}, \cdots, X_p^{(i)})$ be sampled over individuals $i = 1, \cdots, n$ and define

$$n(x_V) \equiv \sum_{i=1}^{n} \mathbb{1}\{X_1^{(i)} = x_1, \cdots, X_p^{(i)} = x_p\}$$

  the number of individuals who have the response pattern x_V. These counts are the sufficient statistics for the multinomial model with log-likelihood

$$\begin{aligned} l(p;n) &= \sum_{x_V} n(x_V) \log p(x_V) \\ &= \sum_{x_V \neq 0_V} n(x_V) \log \frac{p(x_V)}{p(0_V)} + n \log p(0_V) \end{aligned}$$

  where 0_V is the vector of zeros, $p(x_V) \geq 0$, and $\sum_{x_V} p(x_V) = 1$. The multinomial distribution is also an exponential family with

  - Sufficient statistics given by $n(x_V)$.
  - Canonical parameters given by $\log \dfrac{p(x_V)}{p(0_V)}$.
  - Convex cumulant function given by

$$-\log p(0_V) = \log \left(1 + \sum_{x_V \neq 0_V} e^{\theta(x_V)} \right)$$

  Each possibility x_V is called a cell of the table. Given $A \subseteq V$,

$$n(x_A) \equiv \sum_{x_B} n(x_A, x_B)$$

  where $B = V \setminus A$ is called the marginal table.

- The log-linear parameters for $p(x_V) > 0$ are defined by the relation

$$\begin{aligned} \log p(x_V) &= \sum_{A \subseteq V} \lambda_A(x_A) \\ &= \lambda_\varnothing + \lambda_1(x_1) + \cdots + \lambda_V(x_V) \end{aligned}$$

  and the identifiability constraint $\lambda_A(x_A) = 0$ whenever $x_a = 1$ for some $a \in A$.

- Consider a 2×2 contingency table with probabilities π_{ij}. The log-linear parametrisation has

$$\begin{aligned} \log \pi_{11} &= \lambda_\varnothing & \log \pi_{21} &= \lambda_\varnothing + \lambda_X \\ \log \pi_{12} &= \lambda_\varnothing + \lambda_Y & \log \pi_{22} &= \lambda_\varnothing + \lambda_X + \lambda_Y + \lambda_{XY} \end{aligned}$$

We can deduce that
$$\lambda_{XY} = \log \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$$
and $e^{\lambda_{XY}}$ is called the odds ratio between X and Y.

- Let $X_i \sim \mathrm{P}(\mu_i)$ independently, and let $N = \sum_{i=1}^{k} X_i$. Then,
$$N \sim \mathrm{P}(\sum_i \mu_i)$$
$$(X_1, \cdots, X_k)^{\mathrm{T}} \mid N = n \sim \text{Multinom}(n, (\pi_1, \cdots, \pi_k)^{\mathrm{T}})$$
where $\pi_i = \dfrac{\mu_i}{\sum_j \mu_j}$.

- Let $p > 0$ be a discrete distribution on X_V with associated log-linear parameters $\lambda_C, C \subseteq V$. The conditional independence $X_a \perp\!\!\!\perp X_b \mid X_{V\setminus\{a,b\}}$ holds if and only if $\lambda_C = 0$ for all $\{a,b\} \subseteq C \subseteq V$.

Chapter 4 Undirected Graphical Models

- Let V be a finite set. An undirected graph $\mathscr{G}$ is a pair (V,E) where
  - V are the vertices/nodes.
  - $E \subseteq \{\{i,j\} : i,j \in V, i \neq j\}$ is a set of unordered distinct pairs of V called edges.

  We represent graphs by drawing the vertices and then joining pairs of vertices by a line if there is an edge between them.

  - We write $i \sim j$ if $\{i,j\} \in E$, and say they are adjacent in the graph. The vertices adjacent to i are called the neighbours of i, and the set of neighbours is often called the boundary of i and denoted by $\mathrm{bd}_{\mathscr{G}}(i)$.
  - A path in a graph is a sequence of adjacent vertices without repetition. The length of a path is the number of edges in it.
  - Given a subset of vertices $W \subseteq V$, we define the induced subgraph $\mathscr{G}_W$ of $\mathscr{G}$ to be the graph with vertices W, and all edges from $\mathscr{G}$ whose endpoints are contained in W.
  - We say $C \subseteq V$ is complete if $i \sim j$ for every $i,j \in C$. A maximal[2] complete set is called a clique. The set of cliques in a graph is denoted by $\mathscr{C}(\mathscr{G})$.

- Let $A,B,S \subseteq V$. We say that A and B are separated by S in $\mathscr{G}$ ($A \perp_s B \mid S[\mathscr{G}]$) if every path from any $a \in A$ to any $b \in B$ contains at least one vertex in S. A and B are separated by S (where $S \cap A = S \cap B = \varnothing$) iff A and B are separated by $\varnothing$ in $\mathscr{G}_{V\setminus S}$.

[2]Maximal means if one is to add another vertex into the graph, the graph wil no longer be complete. However, graphs in the cliques do not necessarily need to have the same number of vertices.

- Let $\mathscr{G}$ be a graph with vertices V, and let p be a probability distribution over the random variables X_V. We say that p satisfies the pairwise Markov property for $\mathscr{G}$ if

$$i \not\sim j \in \mathscr{G} \implies X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}[p]$$

  We say that p satisfies the global Markov property for $\mathscr{G}$ if for any disjoint sets A, B, S

$$A \perp_s B \mid S \subseteq \mathscr{G} \implies X_A \perp\!\!\!\perp X_B \mid X_S[p]$$

- A distribution p is said to factorises according to graph $\mathscr{G}$ if

$$p(x_V) = \prod_{C \in \mathscr{C}(\mathscr{G})} \psi_C(x_C)$$

  The functions $\psi_C : \mathbb{R}^{|c|} \to \mathbb{R}$ are called potentials.

- If $p(x_V)$ factorises according to $\mathscr{G}$, then p is globally Markov with respect to $\mathscr{G}$.

- (Hammersley-Clifford Theorem). If $p(x_V) > 0$ obeys the pariwise Markove property with respect to $\mathscr{G}$, then p factorises according to $\mathscr{G}$.
  - The followings always hold:

$$\text{factorisation} \Rightarrow \text{global Markov} \Rightarrow \text{pairwise Markov}$$

    The following holds if p is strictly positive:

$$\text{pariwise Markov} \Rightarrow \text{factorisation.}$$

- Given a graph $\mathscr{G}$ with vertices $V = A \cup B \cup S$ where A, B, S are disjoint sets. We say that (A, S, B) consititutes a decomposition of $\mathscr{G}$ if:
  - $\mathscr{G}_S$ is complete;
  - A and B are separated by S in $\mathscr{G}$

  If A and B are both non-empty, we say the decomposition is proper. If not, we say the decomposition is a prime.

- A graph is decomposable if either it is complete or there is a proper decomposition (A, S, B) and $\mathscr{G}_{A \cup S}$, $\mathscr{G}_{B \cup S}$ are decomposable.

- Let $C_1, C_2, \cdots, C_k$ be a collection of subsets. We say that the sequence satisfies the running intersection property (RIP) if $\forall j \geq 2$,

$$C_j \cap \bigcup_{i=1}^{j-1} C_i = C_j \cap C_{\sigma(j)} \qquad \sigma(j) < j$$

- If $C_1, \cdots, C_k$ satisfy the running intersection property, then there is a graph whose cliques are $\mathscr{C} = \{C_1, \cdots, C_k\}$.
- Let $\mathscr{G}$ be an undirected graph. A cycle is a sequence of vertices $< v_1, \cdots, v_k >$ for $k \geq 3$ such that there is a path $v_1 - \cdots - v_k$ and an edge $v_k - v_1$. A chord on a cycle is any edge between two vertices not adjacent on the cycle. A graph is chordal or triangulated if whenever there is a cycle of length greater or equal to 4, it contains a chord.
- Let $\mathscr{G}$ be an undirected graph. The followings are equivalent:
  - $\mathscr{G}$ is decomposable;
  - $\mathscr{G}$ is triangulated;
  - Every minimal separator of $a \not\sim b$ is complete;
  - The cliques of $\mathscr{G}$ satisfy the running intersection property, starting with C.
- A forest is a graph that contains no cycles. If a forest is connected we call it a tree.
- Let $\mathscr{G}$ be a decomposable graph, and let $C_1, \cdots, C_k$ be an ordering of the cliques which satisfies RIP. Define the $j-$th separator set for $j \geq 2$ as

$$S_j \equiv C_j \cap \bigcup_{i=1}^{j-1} C_i = C_j \cap C_{\sigma(j)}$$

  by convention $S_1 = \varnothing$.
- Let $\mathscr{G}$ be a graph with decomposition (A, S, B), and let p be a distribution, then p factorises with respect to $\mathscr{G}$ iff its marginals $p(x_{A\cup S})$ and $p(x_{B\cup S})$ factorise according to $\mathscr{G}_{A\cup S}$ and $\mathscr{G}_{B\cup S}$, and

$$p(x_V) \cdot p_{x_S} = p(x_{A\cup S}) \cdot p(x_{B\cup S})$$

- Let $\mathscr{G}$ be a decomposable graph with cliques $C_1, \cdots, C_k$, then p factorises with respect to $\mathscr{G}$ iff

$$p(x_V) = \prod_{i=1}^{k} p(x_{C_i \setminus S_i} \mid x_{S_i}) = \prod_{i=1}^{k} \frac{p(x_{C_i})}{p(x_{S_i})}$$

- Let $\mathscr{G}$ be an undirected graph, and suppose we have counts $n(x_V)$. Then the MLE $\hat{p}$ under the set of distributions that are Markov to $\mathscr{G}$ is the unique element in which

$$n \cdot \hat{p}(x_C) = n(x_C)$$

  for each clique $C \in \mathscr{C}(\mathscr{G})$.
- The iterative proportional fitting (IPF)/iterative proportional scaling (IPS) algorithm starts with a discrete distribution that satisfies the Markov property for the graph $\mathscr{G}$ (usually pick

uniform distribution), and then iteratively fixes each margin $p(x_C)$ to match the required distribution using the update step:

$$\begin{aligned} p^{(t+1)}(x_V) &= p^{(t)}(x_V) \cdot \frac{\hat{p}(x_C)}{p^{(t)}(x_C)} \\ &= p^{(t)}(x_{V \setminus C} \mid x_C) \cdot \hat{p}(x_C) \end{aligned}$$

The algorithm is: The sequence of distributions in IPF converges to MLE $\hat{p}(x_V)$.

Algorithm 1 IPF algorithm

function IPF(collection of consistent margins $q(x_{C_i})$ for sets $C_1, \cdots, C_k$)
  set $p(x_V)$ to uniform distribution;
  while $\max_i \max_{x_{C_i}} |p(x_{C_i}) - q(x_{C_i})| > \text{tol}$ **do**
    for i in $1, \cdots, k$ **do**
      update $p(x_V)$ to $p(x_{V \setminus C_i} \mid x_{C_i}) \cdot q(x_{C_i})$;
    end for
  end while
  return distribution p with margins $p(x_{C_i}) = q(x_{C_i})$
end function

Chapter 5 Gaussian Graphical Models

- Throughout this course, we assume $\mu = 0$. Let $X_V \sim N_p(\mu, \Sigma)$, and A be a $q \times p$ matrix of full rank q. Then,

  $$AX_V \sim N_q(A\mu, A\Sigma A^{\mathrm{T}})$$

  In particular, for any $U \subseteq V$, we have $X_U \sim N_q(\mu_U, \Sigma_{UU})$.

  The MLEs for multivariate Gaussian distribution are

  $$\hat{\mu} = \bar{X}_v = \frac{1}{n}\sum_{i=1}^{n} X_V^{(i)} \qquad \hat{\Sigma} = W = \frac{1}{n}\sum_{i=1}^{n}\left(X_V^{(i)} - \bar{X}_V\right)^2$$

- $X_A \perp\!\!\!\perp X_B$ iff $\Sigma_{AB} = 0$. $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z$ implies $X \perp\!\!\!\perp Y, Z$ for jointly Gaussian random variables.

- Let $\mathscr{G}$ be a graph with a decomposition (A, S, B), and $X_V \sim N_m(0, \Sigma)$ where $m = |V|$. Then, X_V satisfies the global Markov property with respect to $\mathscr{G}$ only if

  $$\Sigma^{-1} = \{(\Sigma_{A\cup S, A\cup S})^{-1}\}_{A\cup S, A\cup S} + \{(\Sigma_{B\cup S, B\cup S})^{-1}\}_{B\cup S, B\cup S} - \{(\Sigma_{S,S})^{-1}\}_{S,S}$$

  Applying this result to a decomposable graph repeatedly, we see that X_V is Markov with respect to $\mathscr{G}$ iff

  $$\Sigma^{-1} = \sum_{i=1}^{k}\{(\Sigma_{C_i, C_i})^{-1}\} - \sum_{i=2}^{k}\{(\Sigma_{S_i, S_i})^{-1}\}_{S_i, S_i}$$

Note this notation: If M is a matrix whose rows and columns are indexed by $A \subseteq V$, we write $\{M\}_{A,A}$ to indicate the matrix indexed by V (i.e. it has $|V|$ rows and columns) whose $A,A-$entries are M and with zeros elsewhere. For example, if $|V| = 3$, then

$$M = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \qquad \{M\}_{12,12} = \begin{pmatrix} a & b & 0 \\ b & c & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

where 12 is used as an abbreviation for $\{1,2\}$ in the subscript.

- MLE for a decomposable Gaussian graphical model is the unique Σ such that $k_{ij} = 0$ if $i \not\sim j$ and $\sigma_{ij} = w_{ij}$ if $i \sim j$.

Chapter 6 Directed Graphs

- A directed graph $\mathscr{G}$ is a pair (V, D) where
  - V is a set of vertices.
  - $D \subseteq \{(i, j) : i, j \in V, i \neq j\}$ is a set of ordered distinct pairs of V called edges. If $(i, j) \in D$, we write $i \to j$.

  We represent graphs by drawing the vertices and then joining pairs of vertices by a directed line if there is an edge between them.

  - A path in a graph is a sequence of adjacent vertices without repetition. The length of a path is the number of edges in it. A path with length 0 is a single vertex. The path is directed if all the arrows point away from the start.
  - A directed cycle is a directed path from i to $j \neq i$, together with $j \to i$.
  - Graphs that contain no directed cycles are called acyclic, or directed acyclic graphs (DAGs).
  - Note the following concepts:

$$i \to j \quad \begin{cases} i \in \mathrm{pa}_{\mathscr{G}}(j) & i \text{ is a parent of } j \\ j \in \mathrm{ch}_{\mathscr{G}}(i) & j \text{ is a child of } i \end{cases}$$

$$a \to \cdots \to b \text{ or } \mathrm{a} = \mathrm{b} \begin{cases} a \in \mathrm{an}_{\mathscr{G}}(b) & a \text{ is an ancestor of } b \\ b \in \mathrm{de}_{\mathscr{G}}(a) & b \text{ is a descendant of } a \end{cases}$$

  If $w \notin \mathrm{de}_{\mathscr{G}}(v)$, then w is a non-descendant of v:

$$\mathrm{nd}_{\mathscr{G}}(v) = V \setminus \mathrm{de}_{\mathscr{G}}(v)$$

- If the graph is acyclic, we can find a topological ordering (i.e. one in which no vertex comes before any of its parents). Given x_i, define

$$\text{pre}(i) = \{x_1, \cdots, x_{i-1}\}$$

- For any multivariate distribution, we can factorise it as

$$p(x_V) = \prod_{i=1}^{m} p(x_i \mid x_{i,\cdots,x_{i-1}})$$

Let $\mathscr{G}$ be a DAG. We say that $p(x_V)$ factorises according to $\mathscr{G}$ if

$$p(x_V) = \prod_{i=1}^{m} p(x_i \mid x_{\text{pa}_{\mathscr{G}}}(i))$$

Given a topological ordering, we require that

$$p(x_i \mid x_{1,\cdots,x_{i-1}}) = p(x_i \mid x_{\text{pa}_{\mathscr{G}}(i)})$$

which is to say $X_i \perp\!\!\!\perp X_{\text{pre}(i)\setminus\text{pa}(i)} \mid X_{\text{pa}(i)}$. Note that the ordering is arbitrary, let $\text{pre}(i) = \text{nd}_{\mathscr{G}}(i)$ and obtain

$$X_i \perp\!\!\!\perp X_{\text{nd}(i)\setminus\text{pa}(i)} \mid X_{\text{pa}(i)} \qquad \forall i \in V$$

Distributions that have these independences satisfy the local Markov property.

- A set of vertices is ancestral if it contains all of its own ancestors.
- Let $\mathscr{G}$ be a DAG with an ancestral set A. Then, $p(x_V)$ factorises according to $\mathscr{G}$ only if $p(x_A)$ factorises according to $\mathscr{G}_A$.
- A v-structure/unshielded collider is a triple $i \to k \leftarrow j$ where $i \not\sim j$.
- The moral graph for a DAG $\mathscr{G}$ is an undirected graph $\mathscr{G}^m$ such that

$$i \sim j\ [\mathscr{G}^m] \Leftrightarrow \begin{cases} i \sim j\ [\mathscr{G}] \\ i \to k \leftarrow j\ [\mathscr{G}] \end{cases}$$

- If $p(x_V)$ factorises according to a DAG $\mathscr{G}$, then it also factorises according to $\mathscr{G}^m$.
- We say that a probability density $p(x_V)$ satisfies the global Markov property for a DAG $\mathscr{G}$ if wherever $A \perp_s B \mid C$ in $(\mathscr{G}_{\text{an}(A\cup B\cup C)})^m$ we have $X_A \perp\!\!\!\perp X_B \mid X_C$ under p.
- Let $\mathscr{G}$ be a DAG and $p(x_V)$ be a probability density. Then, the followings are equivalent:
  - p factorises according to $\mathscr{G}$;
  - p is globally Markov with respect to $\mathscr{G}$;
  - p is locally Markov with respect to $\mathscr{G}$.

E.g.: Let X_V be a multinomial random vector with probabilities $p(x_V)$, then the log-likelihood if p factorises according to a DAG $\mathscr{G}$ is

$$\begin{aligned}
l(p;n) &= \sum_{x_v \in X_V} n(x_v) \log p(x_v) \\
&= \sum_{x_v \in X_V} n(x_V) \sum_{i=1}^{m} \log p(x_i \mid x_{\mathrm{pa}(i)}) \\
&= \sum_{i=1}^{m} \sum_{x_v \in X_V} n(x_V) \log p(x_i \mid x_{\mathrm{pa}(i)}) \\
&= \sum_{i=1}^{m} \sum_{x_{i \cup \mathrm{pa}(i)} \in X_{\{i\} \cup \mathrm{pa}(i)}} \log p(x_i \mid x_{\mathrm{pa}(i)}) \sum_{x_{V \setminus (\{i\} \cup \mathrm{pa}(i))} \in X_{V \setminus (\{i\} \cup \mathrm{pa}(i))}} n(x_V) \\
&= \sum_{i=1}^{m} \sum_{x_{\{i\} \cup \mathrm{pa}(i)}} \log p(x_i \mid x_{\mathrm{pa}(i)})
\end{aligned}$$

The MLE can then be calculated:

$$\hat{p}(x_i \mid x_{\mathrm{pa}(i)}) = \frac{n(x_i, x_{\mathrm{pa}(i)})}{n(x_{\mathrm{pa}(i)})}$$

For a Bayesian, one may have parameters Θ_i for each $p(x_i \mid x_{\mathrm{pa}(i)})$. If we choose independent priors, then

$$\begin{aligned}
\pi(\theta \mid n) &\propto \pi(\theta) L(\theta \mid n) \\
&= \prod_{i=1}^{m} \pi(\theta_i) f(\theta_i; n(x_i, x_{\mathrm{pa}(i)}))
\end{aligned}$$

This is to say that

$$\theta_i \perp\!\!\!\perp X_{V \setminus (\{i\} \cup \mathrm{pa}(i))}, \theta_{V \setminus \{i\}} \mid X_i, X_{\mathrm{pa}(i)}$$

- For undirected graphs, missing edge induces an independence, hence all graphs give distinct models. If two graphs $\mathscr{G}$ and $\mathscr{G}'$ induce the same statistical model, we say that they are Markov equivalent. E.g.: These models are Markov equivalent:

  - $X \to Z \to Y$:
  $$p(x)p(z \mid x)p(y \mid z) \Leftrightarrow X \perp\!\!\!\perp Y \mid Z$$
  - $X \leftarrow Z \leftarrow Y$.
  - $X \leftarrow Z \to Y$:
  $$p(z)p(x \mid z)p(y \mid z) \Leftrightarrow X \perp\!\!\!\perp Y \mid Z$$
  - $X - Z - Y$:
  $$p(x,y,z) = \psi_{XZ}(x,z) \cdot \psi_{YZ}(y,z) \Leftrightarrow X \perp\!\!\!\perp Y \mid Z$$

- Given a DAG $\mathscr{G}$, we define its skeleton as the undirected graph skel$(\mathscr{G})$ with the same nodes/vertices and the same adjacencies as $\mathscr{G}$.
- Let $\mathscr{G}, \mathscr{G}'$ be graphs with different skeletons (DAG or undirected). Then, $\mathscr{G}$ and $\mathscr{G}'$ are not Markov equivalent.
- Directed graphs $\mathscr{G}$ and $\mathscr{G}'$ are Markov equivalent iff they have the same skeleton and v-structures.
- An undirected graph is Markov equivalent to a directed graph iff it is decomposable.

Chapter 7 Junction Trees and Message Passing

- A connected, undirected graph without any cycles is called a tree, denoted by $\mathscr{T}$. Let $\mathscr{V}$ be vertices contained in the power set of V, that is, each vertex of $\mathscr{T}$ is a subset of V. We say that $\mathscr{T}$ is a junction tree if whenever we have $C_i, C_j \in \mathscr{V}$ with $C_i \cap C_j \neq \varnothing$, there is a (unique) path π in $\mathscr{T}$ from C_i to C_j such that for every vertex C on the path, $C_i \cap C_j \subseteq C$.
- If $\mathscr{T}$ is a junction tree. then its vertices $\mathscr{V}$ can be ordered to satisfy the r.i.p. Conversely, if a collection of sets satisfies the r.i.p., they can be arranged into a junction tree. A tree that does not satisfy r.i.p. is sometimes called a clique tree.
- We will associate each node C in our junction tree with a potential $\psi_C(x_C) \geq 0$, which is a function over the variables in the corresponding set. We say that two potentials ψ_C, ψ_D are consistent if
$$\sum_{x_{C\setminus D}} \psi_C(x_C) = f(x_{C\cap D}) = \sum_{x_{D\setminus C}} \psi_D(x_D)$$
That is, the margins of ψ_C and ψ_D over $C \cap D$ are the same.
- Let $C_1, \cdots, C_k$ satisfy the r.i.p. with separator sets $S_2, \cdots, S_k$, and let
$$p(x_V) = \prod_{i=1}^{k} \frac{\psi_{C_i}(x_{C_i})}{\psi_{S_i}(x_{S_i})}$$
where $S_1 = \varnothing$ and $\psi_\varnothing = 1$ by convention. Then, each $\psi_{C_i}(x_{C_i}) = p(x_{C_i})$ and $\psi_{S_i}(x_{S_i}) = p(x_{S_i})$ iff each pair of potentials is consistent.
- If a graph is not decomposable, then we can triangulate it by adding edges.
- Suppose that two cliques C and D are adjacent in the junction tree, with a separator set $S = C \cap D$. An update from C to D consists of replacing ψ_S and ψ_D with the following:
$$\psi_S'(x_S) = \sum_{x_{C\setminus S}} \psi_C(x_C) \qquad \psi_D'(x_D) = \frac{\psi_S'(x_S)}{\psi_S(x_S)} \psi_D(x_D)$$
This operation is also known as message passing, with the message $\psi_S'(x_S)$ being passed from C to D. We note three important points about this updating step:

- After updating, ψ_C and ψ'_S are consistent.
- If ψ_D and ψ_S are consistent, then so are ψ'_D and ψ'_S.
- The product over all clique potentials

$$\frac{\prod_{C\in\mathscr{C}}\psi_C(x_C)}{\prod_{S\in\mathscr{S}}\psi_S(x_S)}$$

is unchanged: the only altered terms are ψ_D and ψ_S, and by definition of ψ'_D we have

$$\frac{\psi'_D(x_D)}{\psi'_S(x_S)} = \frac{\psi_D(x_D)}{\psi_S(x_S)}$$

Hence, updating preserves the joint distribution and does not upset margins that are already consistent. The junction tree algorithm is a way of updating all the margins such that, when it is complete, they are all consistent.

- Let $\mathscr{T}$ be a tree. Given any node $t \in \mathscr{T}$, we can root the tree at t, and replace it with a directed graph in which all the edges point away from t. The junction tree algorithm involves messages being passed from the edge of the junction tree (the leaves) towards a chosen root (the collection phase), and then being sent away from that root back down to the leaves (the distribution phase). Once these steps are completed, the potentials will all be consistent. This process is also called brief propagation.

Algorithm 2 Collect and distribute steps of the junction tree algorithm

```
function COLLECT(rooted tree 𝒯, potentials ψ_t)
    let 1 < ··· < k be a topological ordering of 𝒯
    for t in k, ··· , 2 do
        send message from ψ_t to ψ_σ(t);
    end for
    return updated potentials ψ_t
end function
function DISTRIBUTE(rooted tree 𝒯, potentials ψ_t)
    let 1 < ··· < k be a topological ordering of 𝒯
    for t in 2, ··· , k do
        send message from ψ_σ(t) to ψ_t;
    end for
    return updated potentials ψ_t
end function
```

- Let $\mathscr{T}$ be a junction tree with potentials $\psi_{C_i}(x_{C_i})$. After running the junction tree algorithm, all pairs of potentials will be consistent.
- In practice, message passing is often done in parallel, and it is not hard to prove that if all potentials update simultaneously, then the potentials will converge to a consistent solution in at most d steps, where d is the width (the length of the longest path) of the tree.

- E.g.: Suppose we have just two tables, ψ_{XY} and ψ_{YZ} arranged in the juntion tree representing a distribution in which $X \perp\!\!\!\perp Z \mid Y$. We can initialise by setting

$$\psi_{XY}(x,y) = p(x \mid y) \qquad \psi_{YZ}(y,z) = p(y,z) \qquad \psi_Y(y) = 1$$

so that $p(x,y,z) = p(y,z) \cdot p(x \mid y) = \dfrac{\psi_{YZ}\psi_{XY}}{\psi_Y}$. Now, we could pick YZ as the root node of our tree, so the collection step consists of replacing

$$\psi_Y'(y) = \sum_x \psi_{XY}(x,y) = \sum_x p(x \mid y) = 1$$

so ψ_Y' and ψ_Y are the same. Hence, the collection step leaves ψ_Y and ψ_{YZ} unchanged. The distribution step consists of

$$\psi_Y''(y) = \sum_z \psi_{YZ}(y,z) = \sum_z p(y,z) = p(y)$$
$$\psi_{XY}'(x,y) = \frac{\psi_Y''(y)}{\psi_Y(y)}\psi_{XY}(x,y) = \frac{p(y)}{1}p(x \mid y) = p(x,y)$$

Hence, after performing both steps, each potential is the marginal distribution corresponding to those variables.

- In junction graphs that are not trees, it is still possible to perform message passing, but convergence is not guranteed. This is known as loopy belief propagation, and is a topic of current research.

- Back to the lung cancer example. Suppose we know a person smokes, then we replace $p(s)$ with $\mathbb{1}_{\{s=1\}}$. Then, we can run a DISTRIBUTE step to obtian other tables conditional on being a smoker. Suppose we have multiple conditions, it is necessary to DISTRIBUTE once for each condition.

Chapter 8 Causal Inference

- A pair $(\mathscr{G}, p)$ is said to be causal if

$$p(x_{V \setminus A} \mid \mathrm{do}(x_A)) = \prod_{i \in V \setminus A} p(x_i \mid x_{\mathrm{pa}(i)}) \qquad \forall A \subseteq V, x_v \in \mathscr{X}_V$$

Here, "do" represents an intervention to set $X_A = x_A$. If we interwene on X, we delete all incoming edges in graph $\mathscr{G}$. Note that it is neither a conditional nor an ordinary marginal distribution.

- Example: Let $Z \to X \to Y, Z \to Y$ be the DAG. We have

$$p(y \mid \mathrm{do}(x)) = \sum_{z \in \mathscr{Z}} p(z)p(y \mid x,z)$$

Note that

$$\begin{aligned} p(y \mid \mathrm{do}(x)) &= \sum_z p(z)p(y \mid z,x) \\ &\neq \sum_z p(z \mid x)p(y \mid z,x) \\ &= p(y \mid x) \end{aligned}$$

This formula is called an adjustment or standardisation or the g-formula.

- Later we usually denote T to be treatment or intervention and Y to be outcome. The following holds:

$$p(y \mid \mathrm{do}(t)) = \sum_{x_{\mathrm{pa}(t)}} p(x_{\mathrm{pa}(t)})p(y \mid t, x_{\mathrm{pa}(t)})$$

- Let $\mathscr{G}$ be a DAG and π a path in $\mathscr{G}$. An internal vertex is any that does not begin or end π. Such a vertex c is a collider if both edges on π contained and point to c (namely, $\to c \leftarrow$). Otherwise it is a non-collider.

- A path form a to b in $\mathscr{G}$ is open conditional on some set $C \subseteq V \setminus \{a,b\}$, if
  - Every collider is in $\mathrm{an}_{\mathscr{G}}(C) = \bigcup_{i \in C} \mathrm{an}_{\mathscr{G}}(i)$ and
  - No non-collider is in C.

  If not, π is blocked given C.

- We say that vertices $a,b \in V$ are d-separated by $C \subseteq V \setminus \{a,b\}$ if every path from a to b is blocked by C. This extends to sets: if every $a \in A$ is d-separated from every $b \in B$ (by C), we say A and B are d-separated by C, denoted by $A \perp_d B \mid C \ [\mathscr{G}]$.

- Let $\mathscr{G}$ be a DAG with disjoint subsets A,B,C as vertices. Then, $A \perp_d B \mid C$ in $\mathscr{G}$ iff $A \perp_s B \mid C$ in $(\mathscr{G}_{\mathrm{an}(A \cup B \cup C)})^m$.

- Any open path from a to b given C is contained in $\mathrm{an}_{\mathscr{G}}(\{a,b\} \cup C)$.

- Given a distribution $p(x_V)$, we say that C is a valid adjustment set for the causal effect of T on Y if

$$p(y \mid \mathrm{do}(t)) = \sum_{x_c \in \mathscr{X}_C} p(x_c)p(y \mid t, x_c)$$

- If we are interested in the total effect of T on Y, then we define the causal nodes as the set of vertices on any causal path (directed path) from T to Y other than T, denoted by $\mathrm{cn}_{\mathscr{G}}(T \to Y)$.

- The forbidden nodes are those that are descendants of any causal node as well as the treatment, denoted by $\mathrm{forb}_{\mathscr{G}}(T \to Y)$.

- We say that C satisfies the generalised adjustment criterion with respect to (t,y) if:
  - C contains no forbidden nodes $(\mathrm{forb}_{\mathscr{G}}(T,Y))$;
  - C blocks all non-causal paths from T to Y.
- Let C satisfy the generalised adjustment criterion with respect to (t,y), then so does $B = C \cap \mathrm{nd}_{\mathscr{G}}(t)$.
- If any $d \in \mathrm{de}_{\mathscr{G}}(t) \cap C$, then either $d \perp_d t \mid C \setminus \mathrm{de}_{\mathscr{G}}(d)$ or $d \perp_d y \mid (\{t\} \cup C) \setminus \mathrm{de}_{\mathscr{G}}(d)$. We refer to $B = C \cap \mathrm{nd}_{\mathscr{G}}(t)$ as a back-door adjustment set.
- If C satisfies the generalised adjustment criterion for (t,y), then

$$\sum_{x_c} p(x_c) p(y \mid t, x_C) = \sum_{x_B} p(x_B) p(y \mid t, x_B)$$

- Suppose C satisfies the generalised adjustment criterion with respect to (t,y), then it is also a valid adjustment set for that pair.
- Let X_V be a multivariate vector of random variables with covariance Σ_{VV}. We denote the coefficients of $X_y = Y$ on X_C by β_{C_y} and $\beta_{t,y \cdot C'}$ where $C \ = C \setminus \{t\}$.
- An increase of one unit in t will increase the expectation of Y by.