

## 2023-2024 Graphical Model Notes

### Chapter 3 Exponential Families and Contingency Tables

- Denote  $X_V \equiv (X_v : v \in V)$ , indexed by  $V = \{1, \dots, p\}$ . Each  $X_v$  takes values in the set  $\mathcal{X}_v$ . 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  $\mathcal{X}$  indexed by  $\theta \in \Theta$ . We say that  $p$  is an **exponential family** if it can be written as

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

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

$$A(\theta) = \log \int \exp \{ \langle \theta, \phi(x) \rangle - C(x) \} dx$$

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

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

$A(\theta)$  and  $-\log p(x; \theta)$  are convex in  $\theta$ , 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  $X_V = (X_1, \dots, X_p)^T \in \mathbb{R}^p$  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  $\mu$  and  $\Sigma$  if the joint density is

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

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

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

we could easily tell that the multivariate Gaussian distribution is an exponential family<sup>1</sup>.

<sup>1</sup>For two matrices  $A$  and  $B$ , we have  $\langle A, B \rangle = \text{tr}(A, B^T)$ .

- 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  $X_V^{(i)} = (X_1^{(i)}, \dots, X_p^{(i)})$  be sampled over individuals  $i = 1, \dots, n$  and define

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

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  $\log \frac{p(x_V)}{p(0_V)}$ .
- 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

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

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

- Consider a  $2 \times 2$  contingency table with probabilities  $\pi_{ij}$ . The log-linear parametrisation has

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

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  $X_i \sim P(\mu_i)$  independently, and let  $N = \sum_{i=1}^k X_i$ . Then,

$$N \sim P\left(\sum_i \mu_i\right)$$

$$(X_1, \dots, X_k)^T \mid N = n \sim \text{Multinom}(n, (\pi_1, \dots, \pi_k)^T)$$

where  $\pi_i = \frac{\mu_i}{\sum_j \mu_j}$ .

- 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**  $\mathcal{G}$  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  $\text{bd}_{\mathcal{G}}(i)$ .
- 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**  $\mathcal{G}_W$  of  $\mathcal{G}$  to be the graph with vertices  $W$ , and all edges from  $\mathcal{G}$  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<sup>2</sup> complete set is called a **clique**. The set of cliques in a graph is denoted by  $\mathcal{C}(\mathcal{G})$ .
- Let  $A, B, S \subseteq V$ . We say that  $A$  and  $B$  are **separated** by  $S$  in  $\mathcal{G}$  ( $A \perp_s B \mid S[\mathcal{G}]$ ) 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  $S \cap A = S \cap B = \emptyset$ ) iff  $A$  and  $B$  are separated by  $\emptyset$  in  $\mathcal{G}_{V \setminus S}$ .

<sup>2</sup>**Maximal** means if one is to add another vertex into the graph, the graph will no longer be complete. However, graphs in the cliques do not necessarily need to have the same number of vertices.

- Let  $\mathcal{G}$  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  $\mathcal{G}$  if

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

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

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

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

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

The functions  $\psi_C : \mathbb{R}^{|C|} \rightarrow \mathbb{R}$  are called **potentials**.

- If  $p(x_V)$  factorises according to  $\mathcal{G}$ , then  $p$  is globally Markov with respect to  $\mathcal{G}$ .
- (**Hammersley-Clifford Theorem**). If  $p(x_V) > 0$  obeys the pairwise Markov property with respect to  $\mathcal{G}$ , then  $p$  factorises according to  $\mathcal{G}$ .

– The followings always hold:

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

The following holds if  $p$  is strictly positive:

$$\text{pairwise Markov} \implies \text{factorisation}.$$

- Given a graph  $\mathcal{G}$  with vertices  $V = A \cup B \cup S$  where  $A, B, S$  are disjoint sets. We say that  $(A, S, B)$  constitutes a **decomposition** of  $\mathcal{G}$  if:

- $\mathcal{G}_S$  is complete;
- $A$  and  $B$  are separated by  $S$  in  $\mathcal{G}$

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  $\mathcal{G}_{A \cup S}, \mathcal{G}_{B \cup S}$  are decomposable.
- Let  $C_1, C_2, \dots, C_k$  be a collection of subsets. We say that the sequence satisfies the **running intersection property (RIP)** if  $\forall j \geq 2$ ,

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

- If  $C_1, \dots, C_k$  satisfy the running intersection property, then there is a graph whose cliques are  $\mathcal{C} = \{C_1, \dots, C_k\}$ .
- Let  $\mathcal{G}$  be an undirected graph. A **cycle** is a sequence of vertices  $\langle v_1, \dots, v_k \rangle$  for  $k \geq 3$  such that there is a path  $v_1 - \dots - v_k$  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  $\mathcal{G}$  be an undirected graph. The followings are equivalent:
  - $\mathcal{G}$  is decomposable;
  - $\mathcal{G}$  is triangulated;
  - Every minimal separator of  $a \not\sim b$  is complete;
  - The cliques of  $\mathcal{G}$  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  $\mathcal{G}$  be a decomposable graph, and let  $C_1, \dots, C_k$  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  $S_1 = \emptyset$ .

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

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

- Let  $\mathcal{G}$  be a decomposable graph with cliques  $C_1, \dots, C_k$ , then  $p$  factorises with respect to  $\mathcal{G}$  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  $\mathcal{G}$  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  $\mathcal{G}$  is the unique element in which

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

for each clique  $C \in \mathcal{C}(\mathcal{G})$ .

- The **iterative proportional fitting (IPF)/iterative proportional scaling (IPS)** algorithm starts with a discrete distribution that satisfies the Markov property for the graph  $\mathcal{G}$  (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  $C_1, \dots, C_k$ )
  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  $1, \dots, k$  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,

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

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

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

- $X_A \perp\!\!\!\perp X_B$  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  $\mathcal{G}$  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  $\mathcal{G}$  only if

$$\Sigma^{-1} = \{(\Sigma_{AUS, AUS})^{-1}\}_{AUS, AUS} + \{(\Sigma_{BUS, BUS})^{-1}\}_{BUS, BUS} - \{(\Sigma_{S, S})^{-1}\}_{S, S}$$

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

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

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

- MLE for a decomposable Gaussian graphical model is the unique  $\Sigma$  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**  $\mathcal{G}$  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  $i \rightarrow j$ .

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  $j \rightarrow i$ .
- Graphs that contain no directed cycles are called **acyclic**, or **directed acyclic graphs (DAGs)**.
- Note the following concepts:

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

$$a \rightarrow \dots \rightarrow b \text{ or } a = b \quad \begin{cases} a \in \text{an}_{\mathcal{G}}(b) & a \text{ is an } \textbf{ancestor} \text{ of } b \\ b \in \text{de}_{\mathcal{G}}(a) & b \text{ is a } \textbf{descendant} \text{ of } a \end{cases}$$

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

$$\text{nd}_{\mathcal{G}}(v) = V \setminus \text{de}_{\mathcal{G}}(v)$$

- 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

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

- For any multivariate distribution, we can factorise it as

$$p(x_V) = \prod_{i=1}^m p(x_i \mid x_{i-1}, \dots, x_1)$$

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

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

Given a topological ordering, we require that

$$p(x_i \mid x_1, \dots, x_{i-1}) = p(x_i \mid x_{\text{pa}_{\mathcal{G}}(i)})$$

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  $\text{pre}(i) = \text{nd}_{\mathcal{G}}(i)$  and obtain

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

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  $\mathcal{G}$  be a DAG with an ancestral set  $A$ . Then,  $p(x_V)$  factorises according to  $\mathcal{G}$  only if  $p(x_A)$  factorises according to  $\mathcal{G}_A$ .
- A **v-structure/unshielded collider** is a triple  $i \rightarrow k \leftarrow j$  where  $i \not\sim j$ .
- The **moral graph** for a DAG  $\mathcal{G}$  is an undirected graph  $\mathcal{G}^m$  such that

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

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



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  $\mathcal{G}$  is

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

The MLE can then be calculated:

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

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

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

This is to say that

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

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

–  $X \rightarrow Z \rightarrow Y$ :

$$p(x)p(z|x)p(y|z) \Leftrightarrow X \perp\!\!\!\perp Y | Z$$

–  $X \leftarrow Z \leftarrow Y$ .

–  $X \leftarrow Z \rightarrow Y$ :

$$p(z)p(x|z)p(y|z) \Leftrightarrow X \perp\!\!\!\perp Y | Z$$

–  $X - Z - Y$ :

$$p(x, y, z) = \psi_{XZ}(x, z) \cdot \psi_{YZ}(y, z) \Leftrightarrow X \perp\!\!\!\perp Y | Z$$

- Given a DAG  $\mathcal{G}$ , we define its **skeleton** as the undirected graph  $\text{skel}(\mathcal{G})$  with the same nodes/vertices and the same adjacencies as  $\mathcal{G}$ .
- Let  $\mathcal{G}, \mathcal{G}'$  be graphs with different skeletons (DAG or undirected). Then,  $\mathcal{G}$  and  $\mathcal{G}'$  are not Markov equivalent.
- Directed graphs  $\mathcal{G}$  and  $\mathcal{G}'$  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  $\mathcal{T}$ . Let  $\mathcal{V}$  be vertices contained in the power set of  $V$ , that is, each vertex of  $\mathcal{T}$  is a subset of  $V$ . We say that  $\mathcal{T}$  is a **junction tree** if whenever we have  $C_i, C_j \in \mathcal{V}$  with  $C_i \cap C_j \neq \emptyset$ , there is a (unique) path  $\pi$  in  $\mathcal{T}$  from  $C_i$  to  $C_j$  such that for every vertex  $C$  on the path,  $C_i \cap C_j \subseteq C$ .
- If  $\mathcal{T}$  is a junction tree, then its vertices  $\mathcal{V}$  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  $\psi_C, \psi_D$  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  $\psi_C$  and  $\psi_D$  over  $C \cap D$  are the same.

- Let  $C_1, \dots, C_k$  satisfy the r.i.p. with separator sets  $S_2, \dots, S_k$ , and let

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

where  $S_1 = \emptyset$  and  $\psi_{\emptyset} = 1$  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  $\psi_S$  and  $\psi_D$  with the following:

$$\psi'_S(x_S) = \sum_{x_{C \setminus S}} \psi_C(x_C) \quad \psi'_D(x_D) = \frac{\psi'_S(x_S)}{\psi_S(x_S)} \psi_D(x_D)$$

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,  $\psi_C$  and  $\psi'_S$  are consistent.
- If  $\psi_D$  and  $\psi_S$  are consistent, then so are  $\psi'_D$  and  $\psi'_S$ .
- The product over all clique potentials

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

is unchanged: the only altered terms are  $\psi_D$  and  $\psi_S$ , and by definition of  $\psi'_D$  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  $\mathcal{T}$  be a tree. Given any node  $t \in \mathcal{T}$ , 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  $\mathcal{T}$ , potentials  $\psi_t$ )

let  $1 < \dots < k$  be a topological ordering of  $\mathcal{T}$

**for**  $t$  in  $k, \dots, 2$  **do**

send message from  $\psi_t$  to  $\psi_{\sigma(t)}$ ;

**end for**

**return** updated potentials  $\psi_t$

**end function**

**function** DISTRIBUTE(rooted tree  $\mathcal{T}$ , potentials  $\psi_t$ )

let  $1 < \dots < k$  be a topological ordering of  $\mathcal{T}$

**for**  $t$  in  $2, \dots, k$  **do**

send message from  $\psi_{\sigma(t)}$  to  $\psi_t$ ;

**end for**

**return** updated potentials  $\psi_t$

**end function**

---

- Let  $\mathcal{T}$  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,  $\psi_{XY}$  and  $\psi_{YZ}$  arranged in the junction tree representing a distribution in which  $X \perp\!\!\!\perp Z \mid Y$ . We can initialise by setting

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

so that  $p(x, y, z) = p(y, z) \cdot p(x \mid y) = \frac{\psi_{YZ}\psi_{XY}}{\psi_Y}$ . 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  $\psi'_Y$  and  $\psi_Y$  are the same. Hence, the collection step leaves  $\psi_Y$  and  $\psi_{YZ}$  unchanged. The distribution step consists of

$$\begin{aligned} \psi''_Y(y) &= \sum_z \psi_{YZ}(y, z) = \sum_z p(y, z) = p(y) \\ \psi'_{XY}(x, y) &= \frac{\psi''_Y(y)}{\psi_Y(y)} \psi_{XY}(x, y) = \frac{p(y)}{1} p(x \mid y) = p(x, y) \end{aligned}$$

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 guaranteed. 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 obtain 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  $(\mathcal{G}, p)$  is said to be **causal** if

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

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

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

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

Note that

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

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:

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

- Let  $\mathcal{G}$  be a DAG and  $\pi$  a path in  $\mathcal{G}$ . An **internal vertex** is any that does not begin or end  $\pi$ . Such a vertex  $c$  is a **collider** if both edges on  $\pi$  contained and point to  $c$  (namely,  $\rightarrow c \leftarrow$ ). Otherwise it is a **non-collider**.
- A path from  $a$  to  $b$  in  $\mathcal{G}$  is **open** conditional on some set  $C \subseteq V \setminus \{a, b\}$ , if
  - Every collider is in  $\text{an}_{\mathcal{G}}(C) = \bigcup_{i \in C} \text{an}_{\mathcal{G}}(i)$  and
  - No non-collider is in  $C$ .

If not,  $\pi$  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  $A \perp_d B \mid C [\mathcal{G}]$ .
- Let  $\mathcal{G}$  be a DAG with disjoint subsets  $A, B, C$  as vertices. Then,  $A \perp_d B \mid C$  in  $\mathcal{G}$  iff  $A \perp_s B \mid C$  in  $(\mathcal{G}_{\text{an}(A \cup B \cup C)})^m$ .
- Any open path from  $a$  to  $b$  given  $C$  is contained in  $\text{an}_{\mathcal{G}}(\{a, b\} \cup C)$ .
- Given a distribution  $p(x_V)$ , we say that  $C$  is a **valid adjustment set** for the causal effect of  $T$  on  $Y$  if

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

- 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  $\text{cn}_{\mathcal{G}}(T \rightarrow Y)$ .
- The **forbidden nodes** are those that are descendants of any causal node as well as the treatment, denoted by  $\text{forb}_{\mathcal{G}}(T \rightarrow Y)$ .

- We say that  $C$  satisfies the **generalised adjustment criterion** with respect to  $(t, y)$  if:
  - $C$  contains no forbidden nodes ( $\text{forb}_{\mathcal{G}}(T, Y)$ );
  - $C$  blocks all non-causal paths from  $T$  to  $Y$ .
- Let  $C$  satisfy the generalised adjustment criterion with respect to  $(t, y)$ , then so does  $B = C \cap \text{nd}_{\mathcal{G}}(t)$ .
- If any  $d \in \text{de}_{\mathcal{G}}(t) \cap C$ , then either  $d \perp_d t \mid C \setminus \text{de}_{\mathcal{G}}(d)$  or  $d \perp_d y \mid (\{t\} \cup C) \setminus \text{de}_{\mathcal{G}}(d)$ . We refer to  $B = C \cap \text{nd}_{\mathcal{G}}(t)$  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  $\Sigma_{VV}$ . We denote the coefficients of  $X_y = Y$  on  $X_C$  by  $\beta_{C_y}$  and  $\beta_{t, y \cdot C'}$  where  $C' = C \setminus \{t\}$ .
- $X_V$  has a distribution Markov with respect to  $\mathcal{G}$  and multivariate Gaussian iff

$$X_i = \sum_{w \in \text{pa}(i)} b_{iw} X_w + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma_i^2), \text{ iid}$$

- If  $(\mathcal{G}, p)$  is causal, then we call these equation **structural**, and  $p$  a **structural equation model**.

$$X_V = (I - B)^{-1} \varepsilon_V \sim N_m(0, \Sigma) \quad \Sigma = (I - B)^{-1} D (I - B)^{-T}$$

where  $D$  is diagonal. Note that we have

$$(I - B)^{-1} = (I + B + B^2 + \dots + B^{m-1})$$

Furthermore,  $(B^2)_{ij}$  is the sum of all directed paths  $j \rightarrow k \rightarrow i$  of length 2,  $(B^l)_{ij}$  is the sum of all directed paths from  $j$  to  $i$  with length  $l$ :

$$(B^2)_{ij} = \sum_{k=1}^m b_{ik} b_{kj}$$

$$(B^l)_{ij} = \sum_{k_1, \dots, k_l} b_{ik_l} b_{k_l k_{l-1}} \dots b_{k_1 j}$$

Hence,  $B^l = 0$  for  $l > m - 1$ .

- Let  $\mathcal{G}$  be a DAG with variables  $V$ . A **trek** from  $i$  to  $j$  with **source**  $k$  is a pair  $(\pi_l, \pi_r)$  where  $\pi_l$  is a directed path from  $k$  to  $i$ , and  $\pi_r$  is a directed path from  $k$  to  $j$ . The two paths are known as the left and right side of the trek. Thus, a trek is essentially a path without colliders, except that we do allow repetition of vertices.
- Given a trek  $\tau = (\pi_l, \pi_r)$  with source  $k$ , define the **trek covariance** as

$$c(\tau) = d_{kk} \prod_{i \rightarrow j \in \pi_l} b_{ji} \prod_{i \rightarrow j \in \pi_r} b_{ji}$$

Note that  $d_{kk} \neq 1$  if  $D \neq I$ .

- Let  $\Sigma = (I - B)^{-1} D (I - B)^{-T}$  be a covariance matrix that is Markov with respect to a DAG  $\mathcal{G}$ . Then,

$$\sigma_{ij} = \sum_{\tau \in \mathcal{T}_{ij}} c(\tau)$$

where  $\mathcal{T}_{ij}$  is the set of treks from  $i$  to  $j$ . This is known as the **trek rule**.

- Let  $\tilde{\beta}_{C_y} = (\tilde{\beta}_{cy \cdot C \setminus \{c\}})_{c \in C}$  be the vector of regression coefficients from regressing  $Y \in \mathbb{R}^n$  on  $X_c \in \mathbb{R}^{n \times q}$  ( $|C| = q$ ). Then

$$\sqrt{n}(\tilde{\beta}_{C_y} - \beta_{C_y}) \xrightarrow{d} N_q(0, (\Sigma_{CC})^{-1} \Sigma_{yy \cdot C})$$

- Suppose  $C, D \subseteq V$  are both valid adjustment sets for  $(T, Y)$ , and let  $C' = C \setminus D$  and  $D' = D \setminus C$ . Then, if  $y \perp_d D' \mid C \cup \{t\}$  and  $t \perp_d C' \mid D$ , we have

$$\frac{\sigma_{yy \cdot tC}}{\sigma_{tt \cdot C}} \leq \frac{\sigma_{yy \cdot tD}}{\sigma_{tt \cdot D}}$$

For good precision on regression, we want the residual variance in  $Y$  to be smaller and residual variance in  $T$  to be larger.

- The **optimal adjustment set** is  $O_{\mathcal{G}}(T \rightarrow Y) = \text{pa}_{\mathcal{G}}(\text{cn}_{\mathcal{G}}(T \rightarrow Y)) \setminus (\text{cn}_{\mathcal{G}}(T \rightarrow Y) \cup \{T\})$ .
- (**Henkel Theorem**) Let  $\mathcal{G}$  be a causal DAG containing  $T$  and  $Y$ . Then,  $O = O_{\mathcal{G}}(T \rightarrow Y)$  satisfies the generalised adjustment criterion with respect to  $(T, Y)$ , and the variance of  $\hat{\beta}_{ty \cdot O}$  is minimal over all such sets. Namely,  $O$  is the optimal adjustment set.