Graph Model Review

maxchenyuling

May 2021

Abstract

This is the review note of Graphical Models, by Yuling Chen. The original notes accredits to the lecture notes of *Grapical Models*, by Prof Robin Evans.

Contents

1 Conditional Independence

1.1 Conditional Independence and Its Properties

Def 2.1: Let X, Y be RV. w/ density p (or mass function). Then,
(i) the **marginal density** for Y is $P(y) = \int_x P(x, y) dx$;
(ii) the **conditional density** for x given Y is $P(x \mid y) \cdot P(y) = P(x, y), \quad \forall x \cdot y$, and;
(iii) X and Y are **independent** if $p(x \mid y) = p(x), \forall x \in X, y \in y, p(y) > 0 \iff p(x, y) = p(x)p(y)$.

Def 2.2: Let X Y be RVs defined on a product space $\mathcal{X} \times \mathcal{Y}$, and Z another RV. Let the joint density be $p(x, y, z)$. Then x is independent of Y conditional on Z, i.e. $(X \perp\!\!\!\perp Y \mid Z \quad [P])$ if: $P(x \mid y, z) = p(x \mid z) \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z} s.t. P(y, z) > 0$.

- If X, Y are marginally independent, write $X \perp\!\!\!\perp Y$.

Ex 2.3 (Markov chain): Let $x_1, x_2, \ldots,$ be a Markov chain, then:

$$\mathbb{P}\left(x_k = x_k \mid X_1 = x_1, \ldots, x_{k-1} = x_{k-1}\right) = \mathbb{P}\left(x_k = x_k \mid x_{k-1} = x_{k-1}\right)$$

i.e. $X_k \perp\!\!\!\perp X_1, X_2 \ldots, X_{k-2}, X_{k-1}[\mathbb{P}]$.

Ex 2.3.1: Suppose $X_v = (x_1 \ldots x_p)^\top$ is a multivariate Gaussian distribution. Then:

$$\begin{aligned} f\left(x_v; \mu, \Sigma\right) &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left(x_v - \mu\right)^\top \Sigma^{-1}\left(x_v - \mu\right)\right) \\ &\implies x_p \mid x_1 = x_1, \ldots, x_{p-1} = x_{p-1} \\ &\quad \sim N\left(\mu_p - \sum_{p,-p}\left(\sum_{-p,-p}\right)^{-1}\left(x_{-p} - \mu_{-p}\right), \sigma_{pp \cdot 1 \ldots p-1}\right) \end{aligned}$$

where $\Sigma_{p,-p}$ is the $p-$th row of the Σ with the $p-$th column removed;
$\Sigma_{-p,-p}$ is the Σ with both $p-$th row and the $p-$th column removed, and;
$\sigma_{aa \cdot B} = \sigma_{aa} - \Sigma_{aB}(\Sigma_{BB})^{-1}\Sigma_{Ba}$.

$\implies x_p \perp\!\!\!\perp X_i \mid X_{V\setminus\{p,i\}}$ iff $\beta_i = \sum_{p,-p}\left(\sum_{-p,p}\right)^{-1} = 0$.

Thm 2.4 (Properties of Conditional Independence): The followings are equivalent:
(i) $p(x \mid y, z) = p(x \mid z)$ for all x, y, z such that $p(y, z) > 0$;
(ii) $p(x, y \mid z) = p(x \mid z) \cdot p(y \mid z)$ for all x, y, z such that $p(z) > 0$;
(iii) $p(x, y, z) = p(y, z) \cdot p(x \mid z)$ for all x, y, z;
(iv) $p(z) \cdot p(x, y, z) = p(x, z) \cdot p(y, z)$ for all x, y, z;
(v) $p(x, y, z) = f(x, z) \cdot g(y, z)$ for some functions f, g and all x, y, z.
Proof of Thm 2.4: (a) $(i) \implies (iii)$: multiply both sides by $P(y, z)$

$$P(x \mid y, z) = P(x \mid z) \implies p(x, y, z) = P(x \mid z)P(y, z)$$

(b) $(iii) \implies (i)$: divided both sides by $P(y, z)$.
(c) $(iii) \implies (v)$: trivial.
(d) $(v) \implies (iii)$:

$$\begin{aligned}
&P(x, y, z) = f(x, z)g(y, z), \text{ integrate over x both sides.}\\
&\qquad \implies p(y, z) = g(y, z)\int_x f(x, z)dx = g(y, z)\tilde{f}(z) \implies g(y, z) = \frac{p(y, z)}{\tilde{f}(z)}(*)\\
&\qquad \implies p(z) = \tilde{f}(z)\int_y g(y, z)dy, \text{ integrate over y both sides.}\\
&\qquad\qquad = \tilde{f}(z)\tilde{g}(z) \implies \text{ if } p(z) > 0, \tilde{f}, \tilde{g} \neq 0\\
&\qquad \implies p(x, y, z) = f(x, z) \cdot \frac{p(y, z)}{\tilde{f}(z)} \quad \text{by}(*)\\
&\qquad \implies p(x \mid y, z) = \frac{f(x, z)}{f(\tilde{z})}
\end{aligned}$$

• Marginal independence has no implication to conditional independence, and vice versa, i.e. $X \perp\!\!\!\perp Y \;\not\Longrightarrow\; X \perp\!\!\!\perp Y \mid Z$ or $X \perp\!\!\!\perp Y \;\not\Longleftarrow\; X \perp\!\!\!\perp Y \mid Z$.

1.2 Graphoid Axioms

Thm 2.6 (Graphoid Axioms):
(i) $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z$ (*symmetry*)
(ii) $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp Y \mid Z$ (*decomposition*)
(iii) $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp W \mid Y, Z$ (*weak union*)
(iv) $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid Y, Z \implies X \perp\!\!\!\perp Y, W \mid Z$ (*contraction*)
(V) If $p(x, t, w, z) > 0$, then $X \perp\!\!\!\perp W \mid Y, Z$ and $X \perp\!\!\!\perp Y \mid W, Z \implies X \perp\!\!\!\perp Y, W \mid Z$ (*intersection*)
Proof of Thm 2.6:
(i) Follows from Thm 2.4.
(ii) $X \perp\!\!\!\perp Y, W \mid Z \implies p(x, y, w, z) = p(x, z) \cdot p(y, w \mid z)$
$\implies p(x, y, z) = p(x, z)\int_w p(y, w \mid z)ds = p(x, z)p(y \mid z)$.
(iii)/(iv) omitted, see PS1.

(v)

$$\begin{aligned}
p(x,y,w,z) &= f(x,w,z)g(y,w,z) \quad \because X \perp\!\!\!\perp Y \mid W,Z \\
&= \tilde{f}(x,y,z)\tilde{g}(y,w,z) \quad \because X \perp\!\!\!\perp W \mid Y,Z \\
&\implies f(x,w,z) = \frac{\tilde{f}(x,y,z)\tilde{g}(y,w,z)}{g(y,w,z)} = a(x,z)b(w,z) \quad \because LHS \perp\!\!\!\perp Y \\
&\implies p(x,y,w,z) = a(x,z)b(w,z)g(y,w,z) = a(x,z)\tilde{\tilde{g}}(y,w,z) \\
&\implies X \perp\!\!\!\perp Y,W \mid Z \quad \text{[EOP]}
\end{aligned}$$

Remark 2.7: By (ii)-(iv), $X \perp\!\!\!\perp W \mid Y,Z$ and $X \perp\!\!\!\perp Y \mid Z \iff X \perp\!\!\!\perp Y,W \mid Z$.

1.3 Functional Conditional Independence

Remark 2.8: Since $\{Y = y\} \equiv \{Y = y, h(Y) = h(y)\}, \forall h$ measurable function, then:
(i) $p(x \mid y,z) = p(x \mid y, h(y), z)$, and hence;
(ii) $X \perp\!\!\!\perp Y \mid Z \implies X \perp\!\!\!\perp h(Y) \mid Z$ and $X \perp\!\!\!\perp Y \mid h(Y), Z$.

Ex 2.9 (Sufficient Statistics): $T \equiv t(x)$ is **sufficient statistic** of θ if: $L(\theta \mid X = x) = f_\theta(x) = g(t(x),\theta) \cdot h(x)$.
• $\pi(\theta \mid x) \propto L(\theta \mid x) \cdot \pi(\theta) = P_\theta(x) \cdot \pi(\theta) = \pi(\theta) f(t(x),\theta) \cdot g(x) \propto \pi(\theta \mid t(x)) \implies \theta \perp\!\!\!\perp X \mid T(x)$

2 Exponential Family and Contingency Table

•$X_V \equiv \{X_v : v \in V\}$ where $V = \{1, \dots, p\}$ is the index set of the nodes.

Def 3.1 (Exponential Family):

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

where:
- ϕ_i: **sufficient statistic**;
- θ_i: **canonical/natural parameter**.
- $A(\theta) = \log \int \exp\{\langle\theta, \phi(x)\rangle - C(x)\}dx$: **cumulant function**;
- $Z(\theta) \equiv \exp(A(\theta))$: **partition function**.

Lemma 3.1 (Gradients of Expo-Family):
(i) $\nabla_\theta A(\theta) = \mathbb{E}_\theta \phi(x)$;
(ii) $\nabla\nabla_\theta^\top A(\theta) = \mathrm{Cov}_\theta\, \phi(x)$;
(iii) A is convex, because $\mathrm{Cov}_\theta\, \phi(x) \geq 0$.
Proof of Lemma 3.1 (i): (else omitted.)

$$
\begin{aligned}
e^{A(\theta)}\frac{\partial}{\partial\theta_i}A(\theta) &= \frac{\partial}{\partial\theta_i}e^{A(\theta)} \\
&= \frac{\partial}{\partial\theta_i}\int \exp\{\langle\theta,\phi(x)\rangle - C(x)\}dx \\
&= \int \frac{\partial}{\partial\theta_i}\exp\{\langle\theta,\phi(x)\rangle - C(x)\}dx \\
&= \int \phi_i(x)\exp\{\langle\theta,\phi(x)\rangle - C(x)\}dx \\
&= e^{A(\theta)}\int \phi_i(x)\exp\{\langle\theta,\phi(x)\rangle - A(\theta) - C(x)\}dx \\
&= e^{A(\theta)}\mathbb{E}_\theta\phi_i(X)
\end{aligned}
$$

Ex 3.2: , omitted see P12 on the notes.

2.1 Properties of Exponential Families

2.1.1 Empirical Moment Matching

We have

$$
\begin{aligned}
\ell(\theta) = \log L(\theta) &= \sum_{x_i}\langle\theta,\phi(x_i)\rangle - nA(\theta) + \text{ const.} \\
&= \left\langle \theta, \sum_{x_i}^{n}\phi(x_i)\right\rangle - nA(\theta) + \text{ const.} \\
&= n\langle\theta,\overline{\phi(x)}\rangle - nA(\theta) + \text{const} \\
\implies \nabla_\theta l(\theta) &= n\overline{\phi(x)} - n\nabla_\theta A(\theta) \overset{\text{set}}{=} 0 \\
\implies \nabla_\theta A(\theta) &= \overline{\phi(x)} = \mathbb{E}[\phi(x)]
\end{aligned}
$$

Hence MLE is given by $\hat{\theta}$ where $\mathbb{E}_{\hat{\theta}}[\phi(x)] = \overline{\phi(x)}$.

2.1.2 Multivariate Gaussian Distribution

$$
\begin{aligned}
f(X_v;\mu,\Sigma) &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(X_v-\mu)^\top\Sigma^{-1}(X_v-\mu)\right), \forall X_v\in\mathbb{R}^p \\
&= \frac{1}{(2\pi)^{p/2}}\exp\left\{-\frac{1}{2}x_V^TKx_V + \mu^TKx_V - \frac{1}{2}\mu^TK\mu + \frac{1}{2}\log|K|\right\} \\
\implies \log L(\theta,\Sigma) = \ell(\theta,\Sigma) &\propto -\frac{1}{2}X_v^\top KX_v + \mu^\top KX_v - \frac{1}{2}\mu^\top K\mu, \text{ with } K\equiv\Sigma^{-1} \\
&= -\frac{1}{2}\operatorname{tr}\left(X_v^\top KX_v\right) + \mu^\top kX_v + \text{Const} \quad \because X_v^TKX_v \text{ is constant} \\
&= -\frac{1}{2}\operatorname{tr}\left(KX_vX_v^\top\right) + \mu^\top KX_v + \text{Const} \quad \because tr(AB) = tr(BA)
\end{aligned}
$$

So, MV Gaussian is an Exponential family, with canonical parameters $\theta = (-K, \eta\equiv\mu^\top K)$ and $\phi(X_v) = (X_vX_v, X_v)$, hence we have:

$2A(\theta) = 2A(K,\eta) = \eta^T K^{-1}\eta + \log \mid K$

$$\implies \nabla_\eta A(\theta) = K^{-1}\eta = \mu = E_\theta\left(X_v\right) = \bar{X}_v$$

$$2\nabla_K A(\theta) = K^{-T}\eta\eta^T K^{-1} + K^{-1} = \Sigma + \mu\mu^T = 2\mathbb{E}_\theta[\frac{1}{2}X_v X_v^T] = \overline{X_v X_v^T}$$

$$\implies \hat{\mu} = \bar{X}_v$$

$$\bar{\Sigma} = \overline{X_v X_v^\top} - \overline{X_v} \cdot \overline{X_v^T}$$

Prop 3.3: Let X_V have a multivariate Gaussian distribution with concentration matrix $K = \Sigma^{-1}$. Then $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}$ if and only if $k_{ij} = 0$, where k_{ij} is the corresponding entry in the concentration matrix.

Proof of Prop 3.3:

The log density is: $\log f\left(x_V\right) = -\frac{1}{2}\left(x_V - \mu\right)^T K\left(x_V - \mu\right) +$ const, where the constant term does not depend on x_V.

The only term involves x_i and x_j is $-k_{ij}(x_i - \mu_i)(x_j - \mu_j)$, hence $k_{ij} = 0$ iff the density has separate terms for x_i and x_j. [EOP].

2.2 Contingency Table

Suppose:

(i) $x_v \equiv (x_0 : v \in V)$ for some set $V = \{1, \ldots, P\}$;

(ii) $x_A \equiv (X_V : v \in A)$ for any $A \subseteq V$;

(iii) $X_v \in \{1, \ldots, d_v\}$.

Counts:

(i) $n\left(x_0\right) = \sum_{i=1}^n \mathbb{I}\left\{x_1^{(i)} = x_1, \ldots, x_p^{(i)} = x_p\right\}$

(ii) $n\left(x_A\right) = \sum_{i=1}^n \mathbb{I}\left\{X_a^{(i)} = x_a : a \in A\right\} = \sum_{X_{V\setminus A}} n\left(X_A, X_{V\setminus A}\right)$ *(marginal table)*.

Loglike:

$$\mathbb{P}\left(X_v^{(i)} = x_0\right) = p\left(x_v\right), \forall x_v \in \{1, \ldots, d_v\}$$

$$\begin{aligned}
\implies P\left(n\left(x_v\right) : x_v \in X_v\right) &= \frac{n!}{\prod_{x \in X_v}\prod_n n\left(x_v\right)!}\prod_{x_v \in X_v}^{\pi} p\left(x_v\right)^{n(x_v)}, \forall p, \sum_{x_v} p\left(x_v\right) = 1 \\
&= \exp\left\{\sum_{x_v} n(x_v) \cdot \log p(x_v) + \text{Const}\right\} \\
&= \exp\left(\sum_{x_v \neq 0_v} \underbrace{n\left(x_v\right)}_{\phi(x_i)} \underbrace{\log\frac{p\left(x_v\right)}{p\left(0_v\right)}}_{\theta(X_v)\in(-\infty,\infty)} + \underbrace{n\log p\left(0_v\right)}_{nA(\theta)} + \text{Const}\right) \implies \text{Exp Family}
\end{aligned}$$

Save of computer memory: Suppose $V = A \cup B \cup S$ and $X_A \perp\!\!\!\perp X_B \mid X_S$. Then:

$$\begin{array}{ll}
p\left(x_V\right) & \to 2^{a+b+s} - 1 \\
= p\left(x_S\right) \cdot p\left(x_A \mid x_S\right) \cdot p\left(x_B \mid x_S\right) & \to \left(2^s - 1\right) + \left(2^{s+a} - 1\right) + \left(2^{s+b} - 1\right) \\
= P\left(x_A, x_S\right) P\left(x_B \mid x_S\right) & \to \left(2^{a+s} - 1\right) + \left(2^b - 1\right) \times 2^s
\end{array}$$

2.3 Log-Linear Model

Def 3.5 (Log-Linear Model): Let $P(x_0) > 0$. then the $\log -$ linear parameters $\lambda_A(X_A) \cdot A \subseteq V$ are:

$$\log P(x_v) = \sum_{A \leq V} \lambda_A(X_4) \text{ subject to } \lambda_A(x_A) = 0 \text{ if } X_a = 1, \forall a \in A \textit{ (identifiability constraint)}$$

Ex 3.5 (Binary case): omitted, see P15 on notes.

Prop 3.6 Let $X_i \sim \text{Poisson}(\mu_i)$ independently, and let $N = \sum_{i=1}^k X_i$. Then,

$$N \sim \text{Poisson}\left(\sum_i \mu_i\right) \quad \text{and} \quad (X_1, \ldots, X_k)^T \mid N = n \sim \text{Multinom}\left(n, (\pi_1, \ldots, \pi_k)^T\right)$$

where $\pi_i = \mu_i / \sum_j \mu_j$

Proof of Prop 3.6: Poisson likelihood is

$$\begin{aligned}
L(\mu_1, \ldots, \mu_k; x_1, \ldots, x_k) &= \prod_{i=1}^k e^{-\mu_i} \mu_i^{x_i} = \frac{k}{\pi_1} e^{-\mu \pi_i} \left(\sum_{j=1}^k \mu_j\right)^{x_i} \pi_i^{x_i}, \quad \because \pi_i = \frac{\mu_i}{\sum_{j=1}^k \mu_j} \\
&= \left(\sum_{j=1}^k \mu_j\right)^{\sum_{i=1}^k x_i} e^{-(\sum_{j=1}^k \mu_j) \sum_{i=1}^k \pi_i} \prod_{i=1}^k \pi_i^{x_i} \\
&= \left(\sum_{j=1}^k \mu_j\right)^N e^{-(\sum_{j=1}^k \mu_j)} \cdot \prod_{i=1}^k \pi_i^{x_i}, \quad \because \sum_i \pi_i = 1 \\
&= \underbrace{L\left(\left(\sum_{j=1}^k \mu_j\right); N\right)}_{\textit{Poisson}} \cdot \underbrace{L(\pi_1, \ldots, \pi_k; x_1 \ldots, x_k \mid N)}_{\text{Conditional Multinomial}} \quad \text{[EOP]}
\end{aligned}$$

Thm 3.7 (Conditional Independence in Log-Linear Model): Let $P > 0$ discrete distribution on X_V with log-linear parameters $\lambda_C, C \subseteq V$. Then,

$$X_a \perp\!\!\!\perp X_b \mid X_{V \setminus \{a,b\}} \quad [P] \iff \lambda_{\{a,b\} \cup C} = 0, \forall C \subseteq V \setminus \{a,b\} \iff \lambda_W = 0, \forall \{a,b\} \subseteq W \subseteq V$$

Proof of Thm 3.7: omitted, see PS.

Corollary 3.7.1: Consider $A \cup B \cup S = V$ with $X_A \perp\!\!\!\perp X_B \mid S$, then by Thm 2.4 (iii), $p(x_S) \cdot p(x_A, x_B, x_S) = p(x_A, x_S) \cdot p(x_B, x_S)$. Hence, $\log p(x_A, x_B, x_S) = \log p(x_A, x_S) + \log p(x_B, x_S) - \log p(x_S)$ Applying log-linear expansion gives:

$$\sum_{W \subseteq V} \lambda_W(x_W) = \sum_{W \subseteq A \cup S} \lambda_W^{AS}(x_W) + \sum_{W \subseteq B \cup S} \lambda_W^{BS}(x_W) - \sum_{W \subseteq S} \lambda_W^S(x_W) \quad (*)$$

By equating the terms, we have:

$$\begin{aligned}
&\lambda_W(x_W) = \lambda_W^{AS}(x_W) && \text{for any } W \subseteq A \cup S \text{ with } W \cap A \neq \emptyset \\
&\lambda_W(x_W) = \lambda_W^{BS}(x_W) && \text{for any } W \subseteq B \cup S \text{ with } W \cap B \neq \emptyset \\
&\lambda_W(x_W) = \lambda_W^{AS}(x_W) + \lambda_W^{BS}(x_W) - \lambda_W^S(x_W) && \text{for any } W \subseteq S
\end{aligned}$$

- Obviously, equation $(*)$ does not include any λ_W^{ABS} term.

3 Undirected Graphical Model

Def 4.1 (Undirected Graph): Let V be a finite set, then an **Undirected Graph** is $\mathcal{G} = \{V, E\}$, where,
•V is the set of **vertex**;
•$E \subseteq \{i, j : i, j \in V, i \neq j\}$ is the **edge** set.

Def 4.2: j is a **neighbor** of i, i.e. $i \sim j$, if i, j are **adjacent** in the graph. The **boundary** of i is the set of neighbors of i, i.e. $\mathrm{bd}_{\mathcal{G}}(i) = \{j : i \sim j\}$.

Def 4.3 (Separation): For $A, B, S \subseteq V$, $A \perp_s B \mid S \quad [\mathcal{G}$.
• $\forall a \in A, b \in B$, the *path* between a and b must include at least one vertex from S.
•$A \perp_s B \mid S \quad [\mathcal{G}] \iff A \perp_s B \mid \emptyset \mid \mathcal{G}_{V \setminus S}$.

Def 4.3.2 (Path): a sequence of adjacent vertices without repetition.

Def 4.3.2 (Induced Subgraph): For a subset $W \subseteq V$, $\mathcal{G}_W$ is the **induced subgraph** of $\mathcal{G}(V, E)$ with vertex $W \subseteq V$ and edges $E_W = \{(i \sim j) \in E : i, j \in W\}$.

3.1 Markov Properties

Def 4.4 (Pairwise Markov Properties): Consider $p(X_v)$ be a distribution over $X_v \in \mathcal{X}_V$. p satisfies PMP if

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

• Whenever an edge is missing in G there is a corresponding conditional independence in p.

Def 4.6 (Glabal Markov Properties): p satisfies GMP if: $\forall$ disjoint set A, B, S,

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

Prop 4.7: GMP $\implies$ PMP.
Proof of Prop 4.7: If $i \not\sim j$ then obviously any path between i and j must have at least one vertex in $V \setminus \{i, j\}$, hence $\{i\} \perp_s \{j\} \mid V \setminus \{i, j\} \quad [\mathcal{G}]$ by Def 4.3. Further by GMP, $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}} \quad [p]$, which is automatically PMP. [EOP]

3.2 Cliques and Factorization

Def 4.8.1 (Completeness): C is **complete** if $i \sim j, \forall i, j \in C$.

Def 4.8.2 (Clique): a maximal complete set.
• $\mathcal{C}(\mathcal{G})$: the set of cliques in a graph $\mathcal{G}$.

Def 4.9 (Factorization): p *factorizes according to graph* $\mathcal{G}$ if

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

for some **potential functions** ψ_C.

Thm 4.10: Factorization $\implies$ GMP.
Proof of Thm 4.10:

Suppose separation $A \perp_s B \mid S \quad [\mathcal{G}]$.
Construct $\tilde{A} = A \cup \mathrm{bd}_{\mathcal{G}_{V\setminus S}}(A)$ the set of vertex that are connected to A by paths in $\mathcal{G}_{V\setminus A}$. Then Construct $\tilde{B} = V \setminus (\tilde{A} \cup S)$. Therefore, we have:
- $B \cap \tilde{A} = \emptyset$;
- $V = \tilde{A} \cup \tilde{B} \cup S$;
- $A \subseteq \tilde{A}$ and $B \subseteq \tilde{B}$;
- no edge between $\tilde{A}$ and $\tilde{B}$.

By the last point, every clique in $\mathcal{G}$ must be either in $\tilde{A} \cup S$ or $\tilde{B} \cup S$, so, let $\mathcal{C}_A(\mathcal{G}) = \{C \in \mathcal{C}(\mathcal{G}) : C \subseteq \tilde{A} \cup S\}$ and $\mathcal{C}_B(\mathcal{G}) = \mathcal{C}(\mathcal{G}) \setminus \mathcal{C}_A(\mathcal{G})$,

$$\begin{aligned} p(X_V) = \prod_{C \in \mathcal{C}} \psi_C(x_C) &= \prod_{C \in \mathcal{C}_A} \psi_C(x_C) \cdot \prod_{C \in \mathcal{C}_B} \psi_C(x_C), \text{ by factorization} \\ &= f\left(x_{\tilde{A}}, x_S\right) \cdot f\left(x_{\tilde{B}}, x_S\right) \implies X_{\tilde{A}} \perp_s X_{\tilde{B}} \mid S \quad [\mathcal{G}] \end{aligned}$$

By Thm 2.6 (ii) (decomposition) and the third point above, we have $X_A \perp_s X_B \mid S \quad [\mathcal{G}]$. [EOP]

Thm 4.11 (Hammersley-Clifford Theorem): If $p(X_V) > 0$ obeys PMP, then p factorizes according to $\mathcal{G}$.

Remark 4.12: Factorization $\implies$ GMP $\implies$ PMP $\overset{p>0}{\implies}$ Factorization.

3.3 Decomposability

Def 4.13 (Decomposition): Consider disjoint sets A, B, S s.t. $A \cup B \cup S = V$, then (A, B, S) is a *decomposition* of $\mathcal{G}$ if: $\mathcal{G}_S$ is complete and $A \perp_s B \mid S \quad [\mathcal{G}]$.
• The decomposition is **proper** if $A \neq \emptyset$ and $B \neq \emptyset$.

Def 4.15 (Decomposability): $\mathcal{G}$ is **decomposible** if either:
(i) $\mathcal{G}$ is itself complete, OR;
(ii) $\exists (A, B, S)$ a proper decomposition, and both $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are decomposible.

Def 4.16 (Running Intersection Property): Consider $\mathcal{C} = \{C : C \subseteq V\}$, $\mathcal{C}$ satisfies RIP if there is an ordering $C_1, \dots, C_k$ s.t. $\forall j = 2, \dots, k, \exists \sigma(j) < j$ with:

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

• Intersection of each set with all the previously seen objects is contained in a single set.

Prop 4.18: If $C_1, \dots, C_k$ satisfies RIP, then $\exists \mathcal{G}$ whose cliques are precisely (the inclusion maximal elements of) $\mathcal{C} = \{C_1, \dots, C_k\}$.

Def 4.19: Consider an undirected graph $\mathcal{G}$,
(i) **Cycle** is a sequence of vertices $\langle v_1, \dots, v_k \rangle$ $(k \geq 3)$ s.t. $\exists$ paths $v_1 \sim v_2 \sim, \dots, \sim v_k$ and an edge $v_k \sim v_1$.
(ii) **Chord** on a cycle is any edge between 2 vertices that are not adjacent on the cycle.
(iii) $\mathcal{G}$ is **chordal/triangulated** if whenever there is cycle of length ≥ 4, it contains a chord.

Thm 4.20: Consider an undirected graph $\mathcal{G}$, the followings are equivalent:
(i) $\mathcal{G}$ is decomposable;

(ii) $\mathcal{G}$ is triangulated;
(iii) every minimal $(a,b)-$separator is complete;
(iv) cliques of $\mathcal{G}$ satisfies RIP.
Proof of Thm 4.20:
(i) $\implies$ (ii): By induction.
Let $p = |V|$ the number of vertices in the graph $\mathcal{G}$. Then if $p \leq 3$, the result is trivial. So only consider $p \geq 4$
If $\mathcal{G}$ is complete, then $\mathcal{G}$ is triangulated, then there is no chordless cycle, then result is trivial.
If $\mathcal{G}$ is NOT complete, then $\exists$ proper decomposition (A, B, S).
$\implies$ $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are both decomposable (by Def 4.15) and have strictly less vertices than $\mathcal{G}$.
$\implies$ $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are triangulated, by induction hypothesis.
$\implies$ Any cycle containing $a \in A$ and $b \in B$ must passes through S twice. Note that S is the separator which is complete, such cycle must contain at least 1 chord connecting the points in S.
$\implies$ By Def 4.19, $\mathcal{G}$ is triangulated.

(ii) $\implies$ (iii): Show contrapositive.
Def 4.20.1 ($(a,b)-$minimal separator): S is the minimal separator of (a,b) if $a \perp_s b \mid S \implies a \not\perp_s b \mid T, \forall T \subseteq S$.
Suppose S is a minimal separator of (a,b) but S is NOT complete. Then $\exists s_1, s_2 : s_1 \not\sim s_2$ and we have a cycle $a \sim \ldots \sim s_1 \sim \ldots \sim b \sim \ldots \sim s_2 \sim \ldots \sim a$.
Let a' be the vertex on the path $a \sim \ldots \sim s_1$ that is closest to s_1 and is adjacent to s_2. Similarly, let b' be the vertex on the path $s_1 \sim \ldots \sim b$ that is closest to s_1 and is adjacent to s_2. Then we have a chordless cycle $a' \sim \ldots \sim s_1 \sim \ldots \sim b \sim s_2 \sim a$ of length ≥ 4. So, $\mathcal{G}$ is not triangulated.

(iii) $\implies$ (iv): By induction
$p = |V| = 1$, the result is trivial.
For $p > 1$, let $a \not\sim b$ with complete minimal separator S. Let $A = \{v \in V : v \not\perp_s a \mid S\}$ and $B = V \backslash (A \cup S)$. Since $A \neq \emptyset$ and $B \neq \emptyset$, (A, B, S) forms a proper decomposition, $A \perp_s B \mid S \quad [\mathcal{G}]$.
By induction hypothesis, cliques of $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ satisfy RIP (because $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ have fewer vertices than $\mathcal{G}$). Taking $(C_1^A, \ldots, C_k^A)$ and $(C_1^B, \ldots, C_k^B)$ as the set of cliques of $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively, the orderings $\mathcal{C}(\mathcal{G}_{A\cup S})$ and $\mathcal{C}(\mathcal{G}_{B\cup S})$ satisfy RIP.
Since $\mathcal{C}(\mathcal{G}) = \mathcal{C}(\mathcal{G}_{A\cup S}) \cup \mathcal{C}(\mathcal{G}_{B\cup S})$, done.

(iv) $\implies$ (i): By induction.
Suppose $(C_1, \ldots, C_k)$ satisfy RIP. If $k = 1$, C_1 is complete, hence decomposible, done.
For $k > 1$, let $H_k = \bigcup_{i<k} C_i$ and $S_k = C_k \cap H_k = C_k \cap C_{\sigma(k)}$, for some $\sigma(k) < k$.
Then there is a proper decomposition $(C_k \backslash S_k, S_k, H_k \backslash S_k)$, because C_k connects all previous vertices via S_k. Now, we have
- $\mathcal{G}_{C_k} = \mathcal{G}_{C_k \backslash S_k \cup S_k}$ is complete so decomposable;
- $\mathcal{G}_{H_k} = \mathcal{G}_{H_k \backslash S_k \cup S_k}$ has $k-1$ cliques satisfying RIP, hence decomposable by induction hypothesis.
Therefore, $\mathcal{G}$ is decomposable, by Def 4.15.

Corollary 4.21.1: Consider $\mathcal{G}$ decomposable and (A, B, S) a proper decomposition, then $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are also decomposable.
Proof of Corollary 4.21.1: $\mathcal{G}$ decomposable $\implies$ $\mathcal{G}$ triangulated (by Thm 4.20), and so is its any subgraphs. [EOP]

Remark 4.21.2: $\mathcal{G}$ triangulated $\iff$ $\mathcal{G}$ decomposable $\implies$ $\mathcal{G}_W (W \subseteq V)$ decomposable $\iff$ $\mathcal{G}_W$

triangulated.

Def 4.22 (Forest): a graph with no cycle.
• Connected forest is a **tree**.

3.4 Separator Sets

Def 4.23.1 (Separator Set): $j-$th Separator Set for $j \geq 2$ is:

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

with $S_1 = \emptyset$.

Lemma 4.23.2: Consider a decomposition (A, B, S) on the undirected graph $\mathcal{G}$, then:
p factorizes according to $\mathcal{G} \iff$ marginals $p(x_{A\cup S})$ and $p(x_{B\cup S})$ factorizes according to $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$, and $p\left(x_V\right) \cdot p\left(x_S\right) = p\left(x_{A\cup S}\right) \cdot p\left(x_{B\cup S}\right)$.
<u>Proof of Lemma 4.23.2</u>:
$(\Longleftarrow)$

$$\begin{aligned} p\left(x_v\right) &= \frac{P\left(x_A, x_S\right) P\left(x_B, x_S\right)}{P\left(x_S\right)} \\ &= \prod_{C \in \mathcal{C}(\mathcal{G}_{A\cup S})} \psi_C\left(x_C\right) \cdot \prod_{D \in \mathcal{C}(\mathcal{G}_{B\cup S})} \psi_D\left(x_D\right) \cdot \frac{1}{P\left(x_S\right)} \\ &= \prod_{C \in \mathcal{C}(\mathcal{G})} \tilde{\psi}(x_C) \quad \because \text{Thm 4.10} \end{aligned}$$

Since (A, B, S) is a decomposition, every clique of $\mathcal{G}$ is either in $\mathcal{G}_{A\cup S}$ or $\mathcal{G}_{B\cup S}$.
$\Longrightarrow$ p factorizes according to $\mathcal{G}$.
$(\Longrightarrow)$ Suppose p factorizes according to $\mathcal{G}$, then p obeys GMP wrt $\mathcal{G}$. So,

$$A \perp_s B \mid S \quad [\mathcal{G}] \overset{GMP}{\Longrightarrow} X_A \perp\!\!\!\perp X_B \mid X_S \quad [p] \overset{Thm:2.4}{\Longrightarrow} p(x_V)p(x_S) = p(x_A, x_S)p(x_B, x_s)$$

Also factorization gives:

$$\begin{aligned} p\left(x_v\right) &= \prod_{C \in d(g)} \psi_c\left(x_c\right) \\ &= \prod_{C \in \mathcal{C}_B(\mathcal{G})} \psi_C\left(x_C\right) \prod_{D \in \mathcal{C}_B(\mathcal{G})} \psi_D\left(x_D\right) \\ \overset{\int dx_A}{\Longrightarrow} P\left(x_B, x_S\right) &= \prod_{D \in \mathcal{C}_B(\mathcal{G})} \psi_D\left(x_D\right) \cdot \int \prod_{c \in \mathcal{C}_A(\mathcal{G})} \psi_C\left(x_C\right) dx_A \\ &= \prod_{D \in \mathcal{C}(\mathcal{G}_{B\cup S})} \tilde{\psi}_D\left(x_D\right) \cdot f\left(x_S\right) = \prod_{C \in \mathcal{C}(\mathcal{G}_{B\cup S})} \hat{\psi}_C(x_C) \end{aligned}$$

Hence, $p(x_B, x_S)$ factorizes wrt the induced subgraph $\mathcal{G}_{B\cup A}$. Similar proof for $p(x_A, x_S)$. [EOP]

Thm 4.24: Let $\mathcal{G}$ be decomposable graph with cliques $C_1, \ldots, C_k$, then p factorizes wrt $\mathcal{G}$ iff:

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

•$p\left(x_{C_i \backslash S_i} \mid x_{S_i}\right)$ is variation independent, so inference over $p(x_V)$ can be based on separate inferences for each $p(x_{C_i})$ individually.

Proof of Thm 4.24:

($\Longleftarrow$) if p factorizes wrt $\mathcal{G}$, then the setting satisfies the factorization property, done.

($\Longrightarrow$) by induction. If $k=1$, result holds trivially. For $k \geq 2$, let $H_k \equiv \left(\bigcup_{i<k} C_i\right) \backslash S_i$. Then we have: $C_k \backslash S_i \perp_s H_k \mid S_k \quad [\mathcal{G}]$, and hence $(C_k \backslash S_i, H_k, S_k)$ is a proper decomposition of the graph $\mathcal{G}$. Note that $\mathcal{G}_{H_k \cup S_k}$ has $k-1$ cliques. By Lemma 4.23.2,

$$p\left(x_{S_k}\right) \cdot p\left(x_V\right)=p\left(x_{C_k}\right) \cdot p\left(x_{H_k}, x_{S_k}\right)=p\left(x_{C_k}\right) \cdot \prod_{i=1}^{k-1} \frac{p\left(x_{C_i}\right)}{p\left(x_{S_i}\right)}$$

because $p(H_k, S_k)$ factorizes $\mathcal{G}_{H_k \cup S_k}$ and by induction hypothesis. [EOP]

3.5 Non-Decomposable Models

Thm 4.25: Let $\mathcal{G}$ be an undirected graph, and suppose we have counts $n\left(x_V\right)$. Then the MLE $\hat{p}$ under the set of distributions that are Markov to $\mathcal{G}$ is the unique element: $\hat{p}\left(x_C\right)=\frac{n\left(x_C\right)}{n}, \forall C \in \mathcal{C}(\mathcal{G})$.

Iterative Proportional Fitting (IPF) algorithm:

Algorithm 1: Iterative Proportional Fitting (IPF) algorithm

Input: a collection of consistent margins $q(x_{C_i})$ for the cliques $C_1, \ldots, C_k$.
Initialize $p(x_V)$ as uniform distribution.
for $t=1, \ldots, T$ **do**
  while $\max_i \max_{C_i} |p^{(t)}(x_{C_i}) - q^{(t)}(x_{C_i})| > tol$ **do**
    for $i=1, \ldots, k$ **do**
      Update $p^{(t+1)}\left(x_V\right)=p^{(t)}\left(x_V\right) \cdot \frac{p\left(x_{C_i}\right)}{p^{(t)}\left(x_{C_i}\right)}=p^{(t)}\left(x_{V \backslash C_i} \mid x_{C_i}\right) \cdot p\left(x_{C_i}\right)$
    end
  end
end
Return: distribution p with margins $p^{(t)}(x_{C_i}) = q^{(t)}(x_{C_i})$.

4 Gaussian Graphical Model

Def 5.0 (Multivariate Gaussian): $X_V \sim N_p(\mu, \Sigma)$ with

$$f\left(x_V\right)=\frac{1}{(2 \pi)^{p / 2}|\Sigma|^{1 / 2}} \exp \left\{-\frac{1}{2}\left(x_V-\mu\right)^T \Sigma^{-1}\left(x_V-\mu\right)\right\}, \quad x_V \in \mathbb{R}^p$$

Prop 5.1: $X_V \sim N_p(\mu, \Sigma)$ and let A be a $q \times p$ matrix of full rank q, then:

$$A X_V \sim N_q\left(A \mu, A \Sigma A^T\right)$$

•$\forall U \subseteq V, X_U \sim N_q(\Sigma_{UU})$.

4.1 Gaussian Graphical Models

•Σ are positive definite, hence by the Hammersley-Clifford Theorem, PMP/GMP/factorization all lead to the same conditional inde- pendence restrictions, and we say that Σ is "Markov wrt $\mathcal{G}$".

•$X_A \perp\!\!\!\perp X_B \iff \Sigma_{AB} = 0$.
•$X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z \not\Longrightarrow X \perp\!\!\!\perp Y, Z$ for jointly Gaussian random variables.

Thm 5.2: $X_V \sim N_p(\mu, \Sigma)$ for positive definite Σ. Then $p(X_V)$ is *Markov wrt* $\mathcal{G}$ iff $k_{ab} = K_{a,b} \equiv (\Sigma)^{-1}_{a,b} = 0, \forall a \not\sim b$ in $\mathcal{G}$.

Lemma 5.3: Consider undirected graph $\mathcal{G}$ with decomposition (A, B, S) and $X_V \sim N_p(0, \Sigma)$, then $p(X_V)$ is *Markov wrt* $\mathcal{G}$ iff

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

and $\Sigma_{A\cup S,A\cup S}$ and $\Sigma_{B\cup S,B\cup S}$ are Markov with respect to $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively.
• With $A \subseteq V$, denote $\{M\}_{A,A}$ as the $|V| \times |V|$ matrix indexed by V, whose $A - A$ entries are M and the rest are zeros.
<u>Proof of Lemma 5.3</u>:
By Lemma 4.23.2, $X_a \perp\!\!\!\perp X_B \mid X_S \implies p(x_V)p(x_S) = p(x_{A\cup S}p(x_{B\cup S})), \forall x_V \in X_V$.
Substituting p with Gaussian distribution and take log, we have:

$$-\frac{1}{2}x_V^T\Sigma^{-1}x_V - \frac{1}{2}x_S^T(\Sigma_{SS})^{-1}x_S = -\frac{1}{2}x_{AS}^T(\Sigma_{AS,AS})^{-1}x_{AS} - \frac{1}{2}x_{BS}^T(\Sigma_{BS,BS})^{-1}x_{BS} + \text{const}$$
$$\overset{(*)}{\Longrightarrow} x_V^T\Sigma^{-1}x_V + x_S^T(\Sigma_{SS})^{-1}x_S = x_{AS}^T(\Sigma_{AS,AS})^{-1}x_{AS} + x_{BS}^T(\Sigma_{BS,BS})^{-1}x_{BS}$$
$$\overset{(**)}{\Longrightarrow} x_V^T\{\Sigma\}^{-1}x_V + x_V^T\left\{(\Sigma_{SS})^{-1}\right\}_{SS}x_V = x_V^T\left\{(\Sigma_{AS,AS})^{-1}\right\}_{AS,AS}x_V + x_V^T\left\{(\Sigma_{BS,BS})^{-1}\right\}_{BS,BS}x_V$$
$$\Longrightarrow \{\Sigma\}^{-1} + \left\{(\Sigma_{SS})^{-1}\right\}_{SS} = \left\{(\Sigma_{AS,AS})^{-1}\right\}_{AS,AS} + \left\{(\Sigma_{BS,BS})^{-1}\right\}_{BS,BS}$$

$(*)$: We can get rid of the constant term because if we set $x_V = 0$, then both the left and right hand side of the equation equal to 0.
$(**)$: Reconstruct the covariance matrices by matching the dimensions, with $\{\Sigma_{CC}\}_{CC}$ is a $|V|\times|V|$ matrix where the $C - C$ entries take the value of the matrix Σ_{CC} and all else entries are 0. [EOP]

Corollary 5.3.1: X_V is Markov wrt $\mathcal{G}$ iff:

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

4.2 Maximum Likelihood Estimation

Def 5.4: Sufficient statistic for Σ is $W \equiv \frac{1}{n}\sum_{i=1}^{n} X_V^{(i)}X_V^{(i)T}$, where $X_V^{(1)}, \ldots, X_V^{(n)} \overset{iid}{\sim} N_p(0, \Sigma)$.
For decomposable graph $\mathcal{G}$ with cliques $C_1, \ldots, C_k$, the MLE is:

$$\left(\hat{\Sigma}^{\mathcal{G}}\right)^{-1} = \sum_{i=1}^{k}\left\{(W_{C_i,C_i})^{-1}\right\}_{C_i,C_i} - \sum_{i=2}^{k}\left\{(W_{S_i,S_i})^{-1}\right\}_{S_i,S_i}$$

5 Directed Graphical Models

Def 6.1 (Directed Graph): A directed graph $\mathcal{G}$ is a pair (V, D), where:
(i) V is a finite set of vertices; and
(ii) $D \equiv \{(v, w) : v \to w, v, w \in V, v \neq w\} \subseteq V \times V$ is a collection of edges, which are ordered pairs of vertices. Loops (i.e. edges of the form (v, v)) are not allowed.

Def 6.1.1:
(i) $v \to w$: v is the **parent** ($v \in \mathrm{pa}_{\mathcal{G}}(w)$) and w is the **child** ($w \in \mathrm{ch}_{\mathcal{G}}(v)$);
(ii) v, w are **adjacent** if $v \to w$ or $w \to v$;
(iii) A **path** in $\mathcal{G}$ is a sequence of distinct vertices such that each adjacent pair in the sequence is adjacent in $\mathcal{G}$;
(iv) The path is **directed** if all the edges point away from the beginning of the path.

Def 6.2: A graph contains a **directed cycle** if there is a directed path from v to w together with an edge $w \to v$.

Def 6.2.1 (Directed Acyclic Graphs): a directed graph with no directed cycle.

Def 6.2.2 (Topological Ordering): an ordering $(1, \ldots, k)$ of the vertices of the graph s.t. $i \in \mathrm{pa}_{\mathcal{G}}(j) \implies i < j$.

Def 6.2.3:
(i) a is an **ancestor** of v ($a \in \mathrm{an}_{\mathcal{G}}(v)$) if either $a = v$ or $\exists$ a directed path $a \to \cdots \to v$;
(ii) b is an **descendant** of v ($b \in \mathrm{an}_{\mathcal{G}}(v)$) if either $b = v$ or $\exists$ a directed path $v \to \cdots \to b$;
(iii) **non-descendant** $\mathrm{nd}_{\mathcal{G}}(v) \equiv V \setminus \mathrm{de}_{\mathcal{G}}(v)$.

5.1 Markov Properties

Def 6.3 (Factorization Property): Let $\mathcal{G}$ be DAG with vertices V. Then $p(x_V)$ factorizes wrt $\mathcal{G}$ if:

$$p(x_V) = \prod_{v \in V} p\left(x_v \mid x_{\mathrm{pa}_{\mathcal{G}}(v)}\right), \quad x_V \in \mathcal{X}_V$$

Def 6.3.1 (Local Markov Property): X_v obeys LMP if:

$$X_v \perp\!\!\!\perp X_{\mathrm{nd}_{\mathcal{G}}(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)} \mid X_{\mathrm{pa}_{\mathcal{G}}(v)} [p]$$

Def 6.3.2 (Ordered Markov Property): X_v obeys OMP if:

$$X_v \perp\!\!\!\perp X_{\mathrm{pre}_{\mathcal{G}}(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)} \mid X_{\mathrm{pa}_{\mathcal{G}}(v)} [p]$$

where $\mathrm{pre}_{\mathcal{G}}(v) = \{i \in V : i < v\}$.
• Under the topological ordering, LMP and OMP are equivalent.

5.2 Ancestrality

Def 6.4.0 (Ancestrality): $A \subseteq V$ is **ancestral** if it contains all its ancestors.

Prop 6.4: Let A be an ancestral set in $\mathcal{G}$. Then $p(x_V)$ factorizes wrt $\mathcal{G}$ iff $p(x_A)$ factorizes wrt $\mathcal{G}_A$, i.e.

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad [p] \iff X_A \perp\!\!\!\perp X_B \mid X_C \quad [p(X_{\mathrm{an}_{\mathcal{G}}(A,B,C)})]$$

Proof of Prop 6.4: omitted, see PS3.

Def 6.5: A **v-structure** is a triple $i \to k \leftarrow j$ such that $i \not\sim j$.

Def 6.5.1: The **moral graph** $\mathcal{G}^m$ of a DAG $\mathcal{G}$ is form from $\mathcal{G}$ by joining any non-adjacent parents and dropping the direction of edges.
• The moral graph removes all the v-structures in a DAG.

Prop 6.6: $p(X_V)$ factorizes wrt DAG $\mathcal{G} \implies p(X_V)$ factorizes wrt undirected graph $\mathcal{G}^m$.

Def 6.7 (Global Markov Property): $p(x_V)$ satisfies GMP wrt DAG $\mathcal{G}$ if:

$$\forall A, B, C \subseteq V : A \perp_s B \mid C \quad [(\mathcal{G}_{\mathrm{an}(A\cup B\cup C)})^m] \implies X_A \perp\!\!\!\perp X_B \mid X_C \quad [p]$$

Thm 6.8 (Completeness of global Markov property): Let $\mathcal{G}$ be a DAG. There exists a probability distribution p s.t.:

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad [p] \iff A \perp_s B \mid C \quad [(\mathcal{G}_{\mathrm{an}(A\cup B\cup C)})^m]$$

• GMP gives all conditional independences that are implied by the DAG model.

Thm 6.9: Let $\mathcal{G}$ be a DAG and p a probability distribution. Then the following are equivalent:
(i) p factorizes according to $\mathcal{G}$;
(ii) p is globally Markov with respect to $\mathcal{G}$;
(iii) p is locally Markov with respect to $\mathcal{G}$.
Proof of Thm 6.9:
(i) $\implies$ (ii): Let $W = \mathrm{an}_{\mathcal{G}}(A \cup B \cup C)$ and suppose $\exists$ a separation $A \perp_s B \mid C \quad [(\mathcal{G}_w)^m]$.
$\implies p(x_w) = \prod_{i \in W} P\left(x_i \mid X_{\mathrm{pa}(i)}\right), \forall x_W$.
By Prop 6.6, $p(x_W)$ also factorizes according to $(\mathcal{G}_W)^m$.
By Thm 4.10, $p(x_W)$ satisfies $X_A \perp\!\!\!\perp X_B \mid X_C \quad [p]$.
(ii) $\implies$ (iii): Moralizing $\mathcal{G}_{\{i\}\cup \mathrm{nd}(i)}$ will not add any edge, hence $i \perp_s \mathrm{nd}(i) \setminus \mathrm{pa}(i) \mid \mathrm{pa}(i) \quad \left[(\mathcal{G}_{\{i\}\cup \mathrm{nd}(i)})^m\right]$.
By GMP, we have $X_i \perp\!\!\!\perp X_{\mathrm{nd}(i)\setminus \mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)} \quad [p]$.
$\implies p$ is locally Markov wrt $\mathcal{G}$.
(iii) $\implies$ (i): GMP $\implies X_i \perp\!\!\!\perp X_{\mathrm{nd}(i)\setminus \mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)} \quad [p]$.
Let $1, \ldots, k$ be a topological ordering, note that $X_i \perp\!\!\!\perp X_{\mathrm{pre}(i)\setminus \mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)} \quad [p], \forall i \in V$.
By definition of Conditional Independence, $p(x_i \mid x_{\mathrm{pre}(i)}) = p(x_i \mid x_{\mathrm{pa}(i)}), \forall x_i \in X_V$, because $\mathrm{pa}(i) \subseteq \mathrm{pre}(i)$ in topological ordering.
$\implies p(x_v) = \prod_{i=1}^{k} p(x_i \mid x_{\mathrm{pre}(i)}) = \prod_{i=1}^{k} p(x_i \mid x_{\mathrm{pa}(i)})$ [EOP]

5.3 Statistical Inference

The **likelihood for a DAG**:

$$\begin{aligned} l(p;n) &= \sum_{x_V} n(x_V)\log p(x_V) \\ &= \sum_{x_V} n(x_V)\sum_{v\in V}\log p\left(x_v \mid x_{\mathrm{pa}(v)}\right) \\ &= \sum_{v\in V}\sum_{x_v, x_{\mathrm{pa}(v)}} n\left(x_v, x_{\mathrm{pa}(v)}\right)\log p\left(x_v \mid x_{\mathrm{pa}(v)}\right) \\ &= \sum_{v\in V}\sum_{\mathrm{pa}(v)}\sum_{x_v} n\left(x_v, x_{\mathrm{pa}(v)}\right)\log p\left(x_v \mid x_{\mathrm{pa}(v)}\right) \end{aligned}$$

Hence the MLE:

$$\hat{p}\left(x_v \mid x_{\mathrm{pa}(v)}\right) = \frac{n\left(x_v, x_{\mathrm{pa}(v)}\right)}{n\left(x_{\mathrm{pa}(v)}\right)} \implies \hat{p}(x_V) \quad = \prod_{v\in V}\hat{p}\left(x_v \mid x_{\mathrm{pa}(v)}\right) = \prod_{v\in V}\frac{n\left(x_v, x_{\mathrm{pa}(v)}\right)}{n\left(x_{\mathrm{pa}(v)}\right)}$$

Suppose each $v \in V$ has a model for $p(x_v|x_{\mathrm{pa}(v)})$, and we have independent prior $\pi(\theta) = \prod_{v\in V}\pi(\theta_v)$, then:

$$\begin{aligned} \pi(\theta \mid x_V) &\propto \pi(\theta)\cdot p(x_V \mid \theta) \\ &= \prod_v \pi(\theta_v)\cdot p\left(x_v \mid x_{\mathrm{pa}(v)}, \theta_v\right) \\ &\implies \theta_i \perp\!\!\!\perp X_{V\setminus(\{i\}\cup\mathrm{pa}(i))}, \theta_{-i} \mid X_i, X_{\mathrm{pa}(i)} \end{aligned}$$

5.4 Markov Equivalence

Def 6.10 (Markov Equivalence): DAGs $\mathcal{G}$ and $\mathcal{G}'$ are **Markov Equivalent** if $\forall p$ Markov wrt $\mathcal{G}$, it is also Markov wrt $\mathcal{G}'$, and vice versa.

Def 6.11 (skeleton): The skeleton of DAG $\mathcal{G} = (V, D)$ is the undirected graph $\mathrm{skel}(\mathcal{G}) = (V, E)$, where $\{i, j\} \in E$ if and only if either $(i, j) \in D$ or $(j, i) \in D$.
• Drop the orientations of edges in $\mathcal{G}$.

Lemma 6.12: $\mathrm{skel}(\mathcal{G}) \neq \mathrm{skel}(\mathcal{G}') \implies \mathcal{G}$ and $\mathcal{G}'$ are not Markov equivalent.
Proof of Lemma 6.12: Suppse we have $i \to j$ in $\mathcal{G}$ but not $\mathcal{G}'$.
Then for $\mathcal{G}$, we have $p(x_v) = p(x_j \mid x_i)\prod_{k\neq j} p(x_k)$. Note that i and j cannot be conditional independent, given any other subset of of $V\setminus\{i,j\}$.
For $\mathcal{G}'$, the LMP implies:

$$\begin{aligned} X_j &\perp\!\!\!\perp X_{\mathrm{nd(j)}\setminus\mathrm{pa}(j))} \mid X_{\mathrm{pa}(j)} \\ X_i &\perp\!\!\!\perp X_{\mathrm{nd(i)}\setminus\mathrm{pa}(i))} \mid X_{\mathrm{pa}(i)} \\ \implies X_i &\perp\!\!\!\perp X_j \mid X_{\mathrm{pa}(j)} \quad \because i \in \mathrm{nd}(j)\setminus\mathrm{pa}(j) \text{ in } \mathcal{G}' \end{aligned}$$

However $p(X_v)$ (for $\mathcal{G}$) does not implies such independence between X_i and X_j. Hence $\mathcal{G}$, $\mathcal{G}'$ not Markov Equiv. [EOP]

Thm 6.13: 2 DAGs $\mathcal{G}, \mathcal{G}'$ are Markov Equivalent iff:
• $\mathrm{skel}(\mathcal{G}) = \mathrm{skel}(\mathcal{G}')$ and;

• v-struc($\mathcal{G}$) = v-struc($\mathcal{G}'$).
Proof of Thm 6.13 ($\Longleftarrow$): proof of ($\Longrightarrow$) is omitted.
If skel($\mathcal{G}$) $\neq$ skel($\mathcal{G}'$), then by Lemma 6.12 $\mathcal{G}, \mathcal{G}'$ are not Markov Equiv.
So only need to show skel($\mathcal{G}$) = skel($\mathcal{G}'$) and v-struc($\mathcal{G}$) $\neq$ v-struc($\mathcal{G}'$) $\Longrightarrow$ Not Markov Equiv.
Suppose WLOG, $\mathcal{G}$ has a v-structure $a \to c \leftarrow b$, which is not contained in $\mathcal{G}'$.
Let p be a distribution in which all variables other than X_a, X_b, X_c are independent to each other, by factorization property, it is:

$$p(x_V) = p(x_c \mid x_a, x_b) \prod_{v \in V \backslash \{c\}} p(x_v)$$

Since skel($\mathcal{G}$) = skel($\mathcal{G}'$), there must exists either $a \to c \to b, a \leftarrow c \to b$, or $a \leftarrow c \leftarrow b$ in $\mathcal{G}'$, i.e. at least one of a or b is the child of c.
Let $A = \text{an}_{\mathcal{G}'}(a, b, c)$. Then claim that $\nexists\ d \in A : a \to d \leftarrow b$ (does not exists d that forms a v-structure with a and b). This is because, as $d \in A$, d is a ancestor of one of (a, b, c). And if $a \to d \leftarrow b$, then d is a descendant of a, b and c, which forms a cycle, which should never happen in a DAG.
Now that a, b does not have common child, there is no edge between a and b in the moral graph $(\mathcal{G}'_A)^m$. So,

$$a \perp_s b \mid A \backslash \{a, b\} \quad [(\mathcal{G}'_A)^m]$$

But $p(x_c|x_a, x_b)$ in p does not factorizes, so p does not factorize wrt $\mathcal{G}'$. Hence $\mathcal{G}, \mathcal{G}'$ not Markov Equiv. [EOP]

Thm 6.14: A DAG $\mathcal{G}$ is Markov Equiv to its undirected (moral) graph iff it has no v-structure.
Proof of Thm 6.14:
($\Longrightarrow$) p factorizes wrt DAG $\mathcal{G}$ implies it factorizes wrt $\mathcal{G}^m$, by Prop 6.6
($\Longleftarrow$) Suppose p is Markov wrt $\mathcal{G}^m$. Let v be a vertex in $\mathcal{G}$ with no child. Then neighbor$_{\mathcal{G}^m}(v) = \text{pa}_{\mathcal{G}}(v)$. So, $\left(v, \text{pa}_{\mathcal{G}}(v), V \backslash (\{v\} \cup \text{pa}_{\mathcal{G}}(v))\right)$ is a proper decomposition in $\mathcal{G}^m$. By Lemma 4.23.2, we have:
(i) $X_v \perp\!\!\!\perp X_{V \backslash (\{v\} \cup \text{pa}_{\mathcal{G}}(v))} \mid X_{\text{pa}_{\mathcal{G}}(v)} \quad [p]$ (hence p satisfies LMP wrt $\mathcal{G}$), and;
(ii) $p(x_{V \backslash \{v\}})$ is Markov wrt $(\mathcal{G}^m)_{V \backslash \{v\}}$.
Since $\mathcal{G}$ has no v-structure, $(\mathcal{G}^m)_{V \backslash \{v\}} = \left(\mathcal{G}_{V \backslash \{v\}}\right)^m \Longrightarrow p(x_{V \backslash \{v\}})$ is Markov wrt $\left(\mathcal{G}_{V \backslash \{v\}}\right)^m$.
Also note $|\left(\mathcal{G}_{V \backslash \{v\}}\right)^m| < |\mathcal{G}^m| \Longrightarrow p(x_{V \backslash \{v\}})$ is Markov wrt $\mathcal{G}_{V \backslash \{v\}}$, by induction hypothesis. [EOP]

Corollary 6.15: A undirected graph is Markov equivalent to a directed graph iff it is decomposable.

6 Junction Trees and Message Passing

• Given a large network of variables, how to efficiently evaluate conditional and marginal probabilities? And how to update our beliefs given new information?

6.1 Junction Trees

• Arrange potential functions to achieve computational convenience.

Def 7.1 (Junction Tree): $\mathcal{T}$ is a junction tree if,
(i) it is a connected, undirected graph without cycles (i.e. it is a tree), and;
(ii) each vertex is a subset of V, i.e. $C_i \subseteq V$, and;
(iii) whenever we have $C_i, C_j \in \mathcal{V}$ with $C_i \cap C_j \neq \emptyset$, there is a (unique) path π in $\mathcal{T}$ from C_i to C_j such that for every vertex C on the path, $C_i \cap C_j \subseteq C$.
• (iii) $\iff$ $\mathcal{T}$ satisfies RIP, i.e. $C_i \cap \bigcup_{j<i} C_j = C_i \cap C_{\sigma(i)}$.

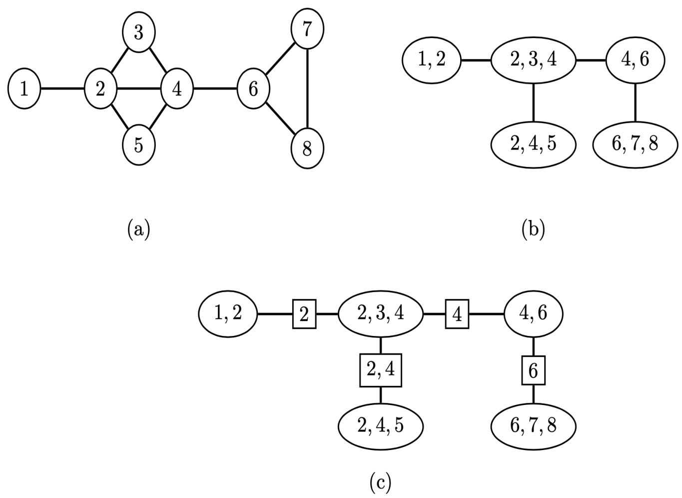


Figure 1: (a) A decomposable graph and (b) a possible junction tree of its cliques. (c) The same junction tree with separator sets explicitly marked.

Prop 7.2: If $\mathcal{T}$ is a junction tree then its vertices $\mathcal{V}$ can be ordered to satisfy the RIP.
• Conversely, if a collection of sets satisfies the RIP they can be arranged into a junction tree.
Proof of Prop 7.2: omitted, see P41-42 on the notes or P44 on the hand notes, with the help of the following corollary.

Corollary 7.2.1: If $C_1, \ldots, C_k$ in a junction tree satisfy RIP, then they satisfy RIP starting with any node C_j.

Def 7.3: For any nodes $C, D \in \mathcal{T}$, the associated potentials ψ_C, ψ_D are **consistent** if the marginal over $C \cap D$ are the same, i.e.

$$\sum_{x_{C\setminus D}} \psi_C(x_C) = f(x_{C\cap D}) = \sum_{x_{D\setminus C}} \psi_D(x_D)$$

Prop 7.4: Let $C_1, \ldots, C_k$ satisfy the RIP with separator sets $S_2, \ldots, S_k$, and let

$$p(x_V) = \prod_{i=1}^{k} \frac{\psi_{C_i}(x_{C_i})}{\psi_{S_i}(x_{S_i})}, \text{ by Thm 4.24}$$

(where $S_1 = \emptyset$ and $\psi_\emptyset = 1$ by convention). Then each pair of potentials is consistent iff:
- $\psi_{C_i}(x_{C_i}) = p(x_{C_i}), \forall i = 1, \ldots, k$ and;
- $\psi_{S_i}(x_{S_i}) = p(x_{S_i}), \forall i = 2, \ldots, k$.
Proof of Prop 7.4: $(\Longleftarrow)$ is trivial as matched potentials are automatically consistent.

$(\Longrightarrow)$ By induction, if $k=1$, done.
For $k>1$, let $R_k = C_k \backslash S_k$ with $S_k = C_k \backslash \bigcup_{i<k} C_i$, so $R_k = C_k \cap \left(\bigcup_{i<k} C_i\right)$ and,

$$\begin{aligned} p\left(x_{V \backslash R_k}\right) &= \sum_{x_{R_k}} p\left(x_V\right) = \prod_{i=1}^{k-1} \frac{\psi_{C_i}\left(x_{C_i}\right)}{\psi_{S_i}\left(x_{S_i}\right)} \times \frac{\sum_{x_{R_k}} \psi_{C_k}\left(x_{C_k}\right)}{\psi_{S_k}\left(x_{S_k}\right)} \\ &= \prod_{i=1}^{k-1} \frac{\psi_{C_i}\left(x_{C_i}\right)}{\psi_{S_i}\left(x_{S_i}\right)} \times \underbrace{\frac{\psi_{C_k \cap S_k}\left(x_{S_k}\right)}{\psi_{S_k}\left(x_{S_k}\right)}}_{=1, \because C_k \cap S_k = S_k} = \prod_{i=1}^{k-1} \frac{\psi_{C_i}\left(x_{C_i}\right)}{\psi_{S_i}\left(x_{S_i}\right)} \end{aligned}$$

Since there are only $k-1$ cliques in $x_{V \backslash R_k}$, by induction hypothesis, $\psi_{C_i}(x_{C_i}) = p(x_{C_i})$ and $\psi_{S_i}(x_{S_i}) = p(x_{S_i}), \forall i<k$.
Further by RIP, $S_k = C_k \cap C_{\sigma(k)}, \sigma(k) < k$. Then by consistency,

$$\psi_{S_k}\left(x_{S_k}\right) = \sum_{x_{C_{\sigma(k)} \backslash S_k}} \psi_{C_{\sigma(k)}}\left(x_{C_{\sigma(k)}}\right) = \sum_{x_{C_{\sigma(k)} \backslash S_k}} p\left(x_{C_{\sigma(k)}}\right) = p\left(x_{S_k}\right)$$

Now that $p\left(x_V\right) = p\left(x_{V \backslash R_k}\right) \frac{\psi_{C_k}\left(x_{C_k}\right)}{\psi_{S_k}\left(x_{S_k}\right)} = p\left(x_{V \backslash R_k}\right) \frac{\psi_{C_k}\left(x_{C_k}\right)}{p\left(x_{S_k}\right)}$,
$\Longrightarrow \frac{\psi_{C_k}\left(x_{C_k}\right)}{p\left(x_{S_k}\right)} = p\left(x_{R_k} \mid x_{V \backslash R_k}\right) = p(x_{R_k} \mid x_{S_k})$, as the LHS only depends on x_{C_k}.
$\Longrightarrow \psi_{C_k}\left(x_{C_k}\right) = p\left(x_{R_k} \mid x_{S_k}\right) \cdot p\left(x_{S_k}\right) = p\left(x_{C_k}\right)$ [EOP]

6.2 Message Passing

- To arrive at the consistent margins.

Algorithm 2: Message Passing from ψ_C to ψ_D

Input: potential function ψ_C, ψ_D, ψ_S with $S = C \cap D$.
Pass the message of $\psi'_S(x_S)$ from C to D involves 2 steps:

$$\text{(a) } \psi'_S\left(x_S\right) = \sum_{x_{C \backslash S}} \psi_C\left(x_C\right)$$

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

Checking Consistency:
(i) After the 2 update steps, ψ_C and ψ'_S are consistent, by (a) step.
(ii) If ψ_D and ψ_S are consistent, then ψ'_D and ψ'_S are also consistent:

$$\sum_{x_{D \backslash S}} \psi'_D\left(x_D\right) \stackrel{(b)}{=} \sum_{x_{D \backslash S}} \frac{\psi'_S\left(x_S\right)}{\psi_S\left(x_S\right)} \psi_D\left(x_D\right) = \frac{\psi'_S\left(x_S\right)}{\psi_S\left(x_S\right)} \underbrace{\sum_{x_{D \backslash S}} \psi_D\left(x_D\right)}_{=\psi_S\left(x_S\right)} = \psi'_S(x_S)$$

(iii) The product over all clique potentials is unchanged $= \frac{\prod_{C \in \mathcal{C}} \psi_C(x_C)}{\prod_{S \in \mathcal{S}} \psi_S(x_S)}$. The only terms that are changed are $\psi_D \to \psi'_D$ and $\psi_S \to \psi'_S$, but the ratio is unchanged by (b) step: $\frac{\psi'_D(x_D)}{\psi'_S(x_S)} = \frac{\psi_D(x_D)}{\psi_S(x_S)}$.

6.3 Junction Tree (Collection-Distribution) Algorithm (JTA)

Algorithm 3: Junction Tree (Collection-Distribution) Algorithm

Collection:
Inputs: rooted tree $\mathcal{T}$, potentials ψ_t.
Let $1 < \ldots < k$ be a topological ordering of $\mathcal{T}$.
for $t = k, \ldots, 2$ **do**
  pass message from ψ_t to $\psi_{\sigma(t)}$
end
Output: updated potentials ψ_t
Distribution:
Inputs: rooted tree $\mathcal{T}$, potentials ψ_t.
Let $1 < \ldots < k$ be a topological ordering of $\mathcal{T}$.
for $t = 2, \ldots, k$ **do**
  pass message from $\psi_{\sigma(t)}$ to ψ_t
end
Output: updated potentials ψ_t

.

Thm 7.5: Let $\mathcal{T}$ be a junction tree with potentials ψ_{C_i}. Then after JTA, all potentials of $\mathcal{T}$ are (pairwise) consistent.
Proof of Thm 7.5: Omitted, see P44 on the notes.

Remark 7.6: If all potentials update simultaneously then the potentials will converge to a consistent solution in at most d steps, where d is the width (i.e. the length of the longest path) of the tree.

Ex 7.7: omitted, see P47 on the notes.

6.4 Directed Graphs and Triangulations

Embed the directed graphical model within a decomposable undirected graph via:
(i) convert to the moral graph;
(ii) *triangulate* the moral graph (by adding chords) until it is decomposable.
• "optimal" triangulation gives the smallest cliques.

Initialization:
Suppose we have a directed graphical model embedded within a decomposable model $C_1, \ldots, C_k$. For each vertex v, the set $\{v\} \cup \mathrm{pa}_{\mathcal{G}}(v)$ is contained within at least one of these cliques. Assigning each vertex arbitrarily to one such clique, let $v(C)$ be the vertices assigned to C. Then set $\psi_C(x_C) = \prod_{v \in v(C)} p\left(x_v \mid x_{\mathrm{pa}(v)}\right)$ and $\psi_S(x_S) = 1$, and we have

$$\prod_{i=1}^{k} \frac{\psi_{C_i}\left(x_{C_i}\right)}{\psi_{S_i}\left(x_{S_i}\right)} = \prod_{v \in V} p\left(x_v \mid x_{\mathrm{pa}(v)}\right) = p\left(x_V\right)$$

• After JTA, the consistent potentials are the marginals for each clique.

6.5 Evidence

• How to incorporate additional information?

Introducing Evidence:
$p(x_{V\setminus E} \mid X_E = x_E^*) = \frac{p(x_{V\setminus E}, x_E^*)}{p(x_E^*)} = \frac{1}{p(x_E^*)} \prod_{i=1}^k \frac{\psi_{C_i}(x_{C_i})}{\psi_{S_i}(x_{S_i})}$
$\implies \psi'_{C_j}(x_{C_j}) \leftarrow \frac{\psi_{C_j}(x_{C_j})}{p(x_E^*)}$, if $E \subseteq C_j$

Prop 7.8: Suppose that potentials Ψ of a junction tree $\mathcal{T}$ with root C is consistent everywhere expect for ψ_C, then running JTA-Distribution$(\mathcal{T}, \Psi)$ starting from C will make everywhere consistent.

Remark 7.9: If we want to introduce multiple evidence in different places, we have to propagate in between by each time running JTa-Distribution step, rooted at which the evidence is introduced.
• The conditional distribution can go wrong if we failed to propagate in between the introductions of the evidences, omitted, see P48 on the nodes for an example.

7 Causal Inference

Def 8.1 (Intervened distribution): $P(Y = y \mid do(X = x))$, the resulting distribution if we intervene the system by setting $X = x$, e.g.
smoking causes cancer but not conversely, then:
- $P(\{\text{ cancer }\} \mid do(\{\text{ smokes }\})) = P(\{\text{ cancer }\} \mid \{\text{ smokes }\})$
- $P(\{\text{ smokes }\} \mid do(\{\text{ cancer }\})) = P(\{\text{ smokes }\})$.

7.1 Intervention

Def 8.2 (Intervention): An **Intervention** on $w \in V$ in a DAG $\mathcal{G}$ with $p(x_V)$ does 2 things:
(i) **graphically**: remove all edges pointing to w, i.e. $v \not\to w, \forall v$;
(ii) **probabilistically**: replace the factorization from $p(x_V) = \prod_{v\in V} p\left(x_v \mid x_{\text{pa}(v)}\right)$ to

$$p\left(x_{V\setminus\{w\}} \mid do\left(x_w\right)\right) = \frac{p\left(x_V\right)}{p\left(x_w \mid x_{\text{pa}(w)}\right)} = \prod_{v\in V\setminus\{w\}} p\left(x_v \mid x_{\text{pa}(v)}\right)$$

• $p(x_w | x_{\text{pa}(w)}) \to \mathbb{I}\{X_w = x_w\}$ i.e. w no longer depend on its parents.

Def 8.2.1: If a graph and its associated probability distribution is **causal**, then an intervention will cause changes both graphically and probabilistically, as in Def 8.2.

Def 8.3 (Confounder): Common cause, e.g. c is a confounder of a and b if $a \leftarrow c \to b$.

Ex 8.3-8.4: omitted, see P51-52 on the notes, and P50 on the hand notes.

7.2 Adjustment Sets and Back-Door Paths

Lemma 8.5: Let $\mathcal{G}$ be a causal DAG. Then the **adjustment formula** gives:

$$p(y \mid do(z)) = \sum_{x_W} \frac{p\left(y, z, x_W\right)}{p\left(z \mid x_{\mathrm{pa}(z)}\right)} = \sum_{x_{\mathrm{pa}(z)}} p\left(y \mid z, x_{\mathrm{pa}(z)}\right) \cdot p\left(x_{\mathrm{pa}(z)}\right)$$

where $X_W = X_V \backslash \{Y, Z\}$.
<u>Proof of Lemma 8.5</u>: Devide X_V into $(Y, Z, X_{\mathrm{pa}(z)}, X_W)$, where X_W is everything remaining. Then:

$$\begin{aligned} p\left(y, x_{\mathrm{pa}(z)}, x_W \mid do(z)\right) &= \frac{p\left(y, z, x_{\mathrm{pa}(z)}, x_W\right)}{p\left(z \mid x_{\mathrm{pa}(z)}\right)}, \text{ by Def 8.2 (ii)} \\ &= p\left(y, x_W \mid z, x_{\mathrm{pa}(z)}\right) \cdot p\left(x_{\mathrm{pa}(z)}\right) \\ \Longrightarrow \quad p(y \mid do(z)) &= \sum_{x_W, x_{\mathrm{pa}(z)}} p\left(y, x_W \mid z, x_{\mathrm{pa}(z)}\right) \cdot p\left(x_{\mathrm{pa}(z)}\right) \\ &= \sum_{x_{\mathrm{pa}(z)}} p\left(x_{\mathrm{pa}(z)}\right) \sum_{x_W} p\left(y, x_W \mid z, x_{\mathrm{pa}(z)}\right) \\ &= \sum_{x_{\mathrm{pa}(z)}} p\left(x_{\mathrm{pa}(z)}\right) p\left(y \mid z, x_{\mathrm{pa}(z)}\right) \quad \text{[EOP]} \end{aligned}$$

Def 8.6 (collider): Let $\mathcal{G}$ be a directed graph and π a path in $\mathcal{G}$. Then **collider** is an internal vertex t on π if the edges adjacent to t meet as $\rightarrow t \leftarrow$.
• Otherwise, t is a **non-collider**: ($\rightarrow t \rightarrow$, $\leftarrow t \leftarrow$, or $\leftarrow t \rightarrow$).

Def 8.7: A path π from a to b is **open** given $C \subseteq V \backslash \{a, b\}$ if:
(i) all colliders on π are in $\mathrm{an}_{\mathcal{G}}(C)$;
(ii) all non-colliders are outside C.

Def 8.7.1: A path is **blocked** by C if it is not open given C.

Def 8.9 (d-separated): Let A, B, C be disjoint sets of vertices in $\mathcal{G}$(C may be empty). We say that A and B are **d-separated** given C (i.e. $A \perp_d B \mid C \quad [\mathcal{G}]$), if every path from $a \in A$ to $b \in B$ is blocked by C.

Thm 8.10: Let $\mathcal{G}$ be a DAG and let A, B, C be disjoint subsets of $\mathcal{G}$. Then A is *d-separated* from B by C in $\mathcal{G}$ iff A is separated from B by C in $\left(\mathcal{G}_{\mathrm{an}(A \cup B \cup C)}\right)^m$, i.e.

$$A \perp_d B \mid C \quad [\mathcal{G}] \iff A \perp_s B \mid C \quad \left[\left(\mathcal{G}_{\mathrm{an}(A \cup B \cup C)}\right)^m\right]$$

<u>Proof of Thm 8.10</u>: omitted, see P54 on the notes and P55 on the hand notes.

7.3 Back Door Adjustments

Def 8.11 (Back-Door Adjustment Set): C is the back-door adjustment set for the order pair (v, w) if:
- no vertex in the C is a descendant of v;

- every path from v to w with an arrow into v (i.e. starting $v \leftarrow \cdots$) is blocked by C.

Thm 8.12: Let C be a *back-door adjustment set* for (v, w), then

$$p\left(x_w \mid do\left(x_v\right)\right)=\sum_{x_C} p\left(x_C\right) \cdot p\left(x_w \mid x_v, x_C\right)$$

i.e. C is a valid adjustment set for the causal distribution.
Proof of Thm 8.12:
First show $v \perp_d C \mid \mathrm{pa}(v)$ (i). Since no vertex in C is a descendant of v, we have that $X_v \perp\!\!\!\perp X_C \mid X_{\mathrm{pa}(v)} \quad [p]$, by LMP. By Thm 8.10, (i) holds.
Then need to show $w \perp_d \mathrm{pa}(v) \mid C \cup \{v\}$ (ii). By contradiction, suppose $\exists$ open path π from w to $t \in \mathrm{pa}(v)$ given $C \cup \{v\}$.
- If it is open given C, then including the edge $t \rightarrow v$ gives an open path from w to v.
- If it is not open given C, then this can only because $\exists$ collider s on π s.t. $s \in \mathrm{an}(v)$ but $s \notin \mathrm{an}(C)$. (Note C does not contain non-colliders.) Hence $\exists$ a directed path from s to v that does not contain any element of C. Then concatenate this directed path with the proportion of π from w to s.
Both ways give an open path from w to v given C, which contradicts that C is a valid back-door adjustment set.
Now, (ii) joint with GMP gives $X_w \perp\!\!\!\perp X_{\mathrm{pa}(v)} \mid X_C, X_v$, then:

$$\begin{aligned}
p\left(x_w \mid do\left(x_v\right)\right) &= \sum_{x_{\mathrm{pa}(v)}} p\left(x_{\mathrm{pa}(v)}\right) \cdot p\left(x_w \mid x_v, x_{\mathrm{pa}(v)}\right) \\
&= \sum_{x_{\mathrm{pa}(v)}} p\left(x_{\mathrm{pa}(v)}\right) \sum_{x_C} p\left(x_w, x_C \mid x_v, x_{\mathrm{pa}(v)}\right) \\
&= \sum_{x_{\mathrm{pa}(v)}} p\left(x_{\mathrm{pa}(v)}\right) \sum_{x_C} p\left(x_w \mid x_C, x_v, x_{\mathrm{pa}(v)}\right) \cdot p\left(x_C \mid x_v, x_{\mathrm{pa}(v)}\right) \\
&= \sum_{x_{\mathrm{pa}(v)}} p\left(x_{\mathrm{pa}(v)}\right) \sum_{x_C} p\left(x_w \mid x_C, x_v\right) \cdot p\left(x_C \mid x_{\mathrm{pa}(v)}\right) \\
&= \sum_{x_C} p\left(x_w \mid x_C, x_v\right) \sum_{\mathrm{pa}(v)} p\left(x_{\mathrm{pa}(v)}\right) \cdot p\left(x_C \mid x_{\mathrm{pa}(v)}\right) \\
&= \sum_{x_C} p\left(x_C\right) \cdot p\left(x_w \mid x_v, x_C\right) \quad \text{[EOP]}
\end{aligned}$$

Prop 8.13: $\mathrm{pa}(v)$ is always a back-door adjustment set.
Proof of Prop 8.13: By Def 8.11, every back-door path starts with $v \leftarrow t \cdots$, where $t \in \mathrm{pa}_{\mathcal{G}}(v)$ is a non-collider on the path. Hence the paths are blocked. [EOP]

• If a variable does not have parent, then ordinary conditional distribution is the same as the causal distribution.

7.4 Gaussian Causal Models

$$
\begin{aligned}
\mathbb{E}[Y \mid do(z)] &= \sum_{x_C} p\left(x_C\right) \cdot \mathbb{E}\left[Y \mid z, x_C\right] \\
&= \int_{\mathcal{X}_C} p\left(x_C\right) \cdot \left(\beta_0 + \beta_z z + \sum_{c \in C} \alpha_c x_c\right) dx_C, \text{ by simple linear model.} \\
&= \beta_0 + \beta_z z + \sum_{c \in C} \alpha_c \mathbb{E} X_c \\
&= \beta_0 + \beta_z z \quad \because \text{ X are standardized s.t. mean} = 0.
\end{aligned}
$$

• β_z is the same for all X_C, hence we can forget the averaging in the adjustment formula and just look at a suitable regression to obtain the causal effect.

7.5 Structural Equation Models

Def 8.14: $(\mathcal{G}, p)$ is a **structural equation model** if $(\mathcal{G}, p)$ is causal and p is a multivariate Gaussian distribution.

Prop 8.14.1: $X_V \sim N_p(0, \Sigma)$ is Markov wrt DAG $\mathcal{G}$ iff

$$
\begin{aligned}
& X_i = \sum_{j \in \mathrm{pa}_{\mathcal{G}}(i)} \beta_{ij} X_j + \varepsilon_i, \quad \epsilon_i \sim N(0, d_{ii}), \quad \forall i \in V \\
\Longrightarrow & \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ \beta_{21} & 0 & 0 & \dots & 0 \\ \beta_{31} & \beta_{32} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{p-1} & 0 \end{pmatrix}_{p \times p} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix} \\
\Longrightarrow & X = BX + \epsilon, \quad \epsilon \sim N_p(\mathbf{0}, D), \text{ with } \operatorname{diag}(D) = (d_{ii})_{i=1}^p \text{ a diagonal matrix} \\
\Longrightarrow & X = (I - B)^{-1} \epsilon \Longrightarrow Cov(X) = (I - B)^{-1} D (I - B)^{-T}
\end{aligned}
$$

• Note that B has the following features:
(i) lower triangular and;
(ii) $\beta_{ij} \neq 0 \iff \exists (j \to i) \in \mathcal{G}$;
(iii) nilpotent $\Longrightarrow (I - B)^{-1} = I + B + B^2 + \cdots + B^{p-1}$.
(iv) $\left(B^2\right)_{ij} = \sum_k \beta_{ik} \beta_{kj}$ with $\beta_{ik} \beta_{kj} \neq 0 \iff (j \to k \to i) \in \mathcal{G}$, and $\left(B^3\right)_{ij} = \sum_k \sum_l \beta_{ik} \beta_{kl} \beta_{lj}$ with $\beta_{ik} \beta_{kl} \beta_{lj} \neq 0 \iff (j \to l \to k \to i) \in \mathcal{G}$;
(v) $B^p = 0$ because the max length of the paths in $\mathcal{G}$ is at most $p - 1$.

7.6 Trek Rule

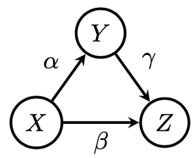
Def 8.16 (Trek): A **trek** from i to j with **source** k is a pair of paths, (π_l, π_r), where
(i) **left trek** π_l is a directed path from k to i, and;
(ii) **right trek** π_r is a directed path from k to j.
• A trek is a path without colliders, but may have repetition of vertices.
• A single vertex can be a trek from itself to itself with source itself, i.e. it is possible to have $k = i$ or $k = j$ or both. (see P58 on the notes for examples.)

Def 8.18 (Trek Covariance): given a trek (π_l, π_r) with source k, the **trek covariance** is:

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

where d_{kk} is the variance of the error term corresponding to the vertex k.

Ex 8.19: with the directed graph with edge coefficients,



Treks from Z to Z:

$$\begin{array}{lll} Z & Z \leftarrow Y \rightarrow Z & Z \leftarrow X \rightarrow Z \\ Z \leftarrow Y \leftarrow X \rightarrow Z & Z \leftarrow X \rightarrow Y \rightarrow Z & Z \leftarrow Y \leftarrow X \rightarrow Y \rightarrow Z \end{array}$$

Trek coefficients:

$$\begin{array}{cc} c(Z) = 1 & c(Z \leftarrow Y) = \gamma \\ c(Z \leftarrow X) = \beta & c(Z \leftarrow X \rightarrow Y \rightarrow Z) = \beta\alpha\gamma \end{array}$$

$\implies \sigma^2_{ZZ} = 1 + \gamma^2 + \beta^2 + 2\alpha\beta\gamma + \alpha^2\gamma^2$

Thm 8.20 (Trek Rule): 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.
Proof of Thm 8.20: By induction.
For $p = 1, \sigma_{11} = 1 \cdot d_{11} = d_{11}$, done.
For $p > 1$, assume the result holds for $|V| < p$, hence it holds on any ancestral subgraphs. Suppose $p \in V$ is a vertex with no child, and X_p is the associated random variable. Then, $X_p = \sum_{j \in \mathrm{pa}_{\mathcal{G}}(p)} b_{pj} X_j + \varepsilon_p$, where $\epsilon_p \perp\!\!\!\perp X_1, \ldots, X_{p-1}$. So, for $i < p$, we have:

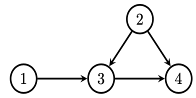
$$\mathrm{Cov}\left(X_i, X_p\right) = \sum_{j \in \mathrm{pa}_{\mathcal{G}}(p)} b_{pj} \,\mathrm{Cov}\left(X_i, X_j\right)$$

with $\mathrm{Cov}\left(X_i, X_j\right) = \sum_{\tau \in \mathcal{T}_{ij}} c(\tau), \forall i, j < p$ by the induction hypothesis.
So, any trek from i to p must consists of a trek from i to j where $j \in \mathrm{pa}_{\mathcal{G}}(p)$, i.e. $j \to p$.
If $i = p$, then we include an extra covariance term of X_p and the corresponding error ϵ_p:

$$\mathrm{Cov}\left(X_p, X_p\right) = \sum_{j \in \mathrm{pa}_{\mathcal{G}}(p)} b_{pj} \,\mathrm{Cov}\left(X_p, X_j\right) + \underbrace{\mathrm{Cov}\left(X_p, \varepsilon_p\right)}_{= Var(\epsilon_p) = d_{pp}}$$

where the first term corresponds to the trek covariance for treks with lengths ≥ 1, and the last term corresponds to the trek covariance for the trek of length 0. [EOP]

Ex 8.21: take the following graph,



Treks from 3 to 3 and the corresponding trek covariance:

$$\begin{array}{lll} 3 & 3 \leftarrow 2 \rightarrow 3 & 3 \leftarrow 1 \rightarrow 3 \\ d_{33} & d_{22} b_{23}^2 & d_{11} b_{13}^2 \end{array}$$

$$\implies \operatorname{Var}(X_3) = \sigma_{33} = d_{33} + d_{22} b_{23}^2 + d_{11} b_{13}^2$$

Treks from 3 to 4 and the corresponding trek covariance:

$$\begin{array}{llll} 3 \rightarrow 4 & 3 \leftarrow 2 \rightarrow 4 & 3 \leftarrow 2 \rightarrow 3 \rightarrow 4 & 3 \leftarrow 1 \rightarrow 3 \rightarrow 4 \\ d_{33} b_{34} & d_{22} b_{23} b_{24} & d_{22} b_{23}^2 b_{34} & d_{11} b_{13}^2 b_{34} \end{array}$$

$$\begin{aligned} \implies \operatorname{Cov}(X_3, X_4) = \sigma_{34} &= d_{33} b_{34} + d_{22} b_{23} b_{24} + d_{22} b_{23}^2 b_{34} + d_{11} b_{13}^2 b_{34} \\ &= \left(d_{33} + d_{22} b_{23}^2 + d_{11} b_{13}^2\right) b_{34} + d_{22} b_{23} b_{24} \\ &= \operatorname{Var}(X_3)\, b_{34} + \operatorname{Var}(X_2)\, b_{23} b_{24} \end{aligned}$$