# Auxiliary notes on Graphical Models

Daniel Lin; Professor: Robin Evans

December 14, 2023

This note aims to explain the more abstract concepts in the course SC6 Graphical Models and supply some useful intuitions to enhance your understanding. Please use it in line with the lecture notes.

## Contents

**1 Introduction 1**

1.1 Simple Applications of CI . . . 4

1.2 Exponential Family . . . 4

1.2.1 Finding MLE by moment-matching . . . 4

1.2.2 Multivariate Normal Distribution as Exponential Family . . . 5

1.3 Contingency Tables . . . 5

1.3.1 Log-linear Models . . . 6

**2 Undirected Graphs 7**

2.1 Decomposability . . . 8

**3 Directed Graphs 15**

3.1 Markov Equivalence . . . 18

**4 Inference on Graphs 22**

4.1 Message Passing . . . 23

4.2 Junction Trees . . . 25

**5 Causal Inference 28**

5.1 Observational Study and Rubin's model . . . 31

5.2 Causal Diagrams and Pearl's model . . . 32

5.3 Intervention and Adjustments . . . 36

5.4 Systematic Bias . . . 38

5.4.1 Confounding . . . 39

5.4.2 Single World Intervention Graph . . . 42

5.4.3 Selection Bias . . . 45

5.5 Adjustment Criterions . . . 47

5.6 Parametric Models . . . 52

5.6.1 Linear Models . . . 52

5.6.2 General Models . . . 54

5.7 Gaussian Structural Equation Model . . . 54

5.8 Efficient Adjustment . . . 58

5.8.1 Forbidden Projection . . . 62

**6 Ending 62**

## 1 Introduction

Suppose we know $Z$, one is interested in whether investigating $Y$ provides information about $X$. e.g. given that the stock market in the wine industry dropped yesterday, would the stock price of company B be relevant to company A, in which I hold shares? Knowing such relevance in any network(relationships, social networks, internet, website connections etc.) saves computational time. e.g. I can break down the stock market with thousands of companies into groups of companies, with no relevance between groups. So studying the dozens of companies in a group with your interested company is enough. With the probabilistic approach, conditional dependence/independence describes the extent of relevance.

**Definition 1** (Conditional Probability)**.** In some textbooks, this is defined as

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

i.e. the probability of $x$ happening when $y$ happens.
But this breaks down if $p(y) = 0$. The definition that requires no additional assumption is $p(x|y)$ is the unique function that makes

$$p(x,y) \equiv p(x|y)p(y)$$

**Definition 2** (Conditional Independence)**.** Recall the independence of two events being defined by $p(x|y) = p(x)$ ($p(y) \neq 0$) or $p(x,y) = p(x)p(y)$. Conditional independence assumes all these probabilities are based on $z$, i.e. adding condition $z$ in every probability.

$$p(x|y,z) = p(x|z)$$

whenever $p(y,z) > 0$. With symbols: $X \perp Y \mid Z$. In this note, I may abbreviate conditional independence as CI.

**Theorem 1.1** (Equivalent Definitions of Conditional Independence)**.** *The following are equivalent:*

- *(i)* $p(x|y,z) = p(x|z)$ *whenever* $p(y,z) > 0$
- *(ii)* $p(x,y|z) = p(x|z)p(y|z)$ *whenever* $p(z) > 0$
- *(v)* $p(x,y,z) = f(x,z)g(y,z)$ *for some functions* $f, g$

*The first one is the traditional definition saying if $z$ is given, the probability distribution of $x$ does not depend on $y$. The second is the conditional (condition on $z$) version of $p(x,y) = p(x)p(y)$ which is easier to manipulate compared to the first one. The third one shares the idea of "decomposition" with the second one, but the components can be arbitrary functions instead of probability density functions (integrate to 1). Therefore, the third one is more practical.*

*Proof.* The equivalence of the first two can be proved similarly to independence without condition.

**(ii) ⇒ (v)**
By multiplying $p(z)$ to both sides of (ii), LHS becomes $p(x,y|z)p(z) = p(x,y,z)$. And setting $f(x,z) := p(x|z)p(z) = p(x,z)$, $g(y,z) := p(y|z)$ finishes the proof. Note you can also set $f(x,z) := p(x|z)$, $g(y,z) := p(y|z)p(z) = p(y,z)$.

**(v) ⇒ (ii)**
(ii) is equivalent to (iii) $p(x,y,z) = p(y,z)p(x|z)$ as shown above. So prove (v) ⇒ (iii) suffices.

If either $p(y,z) = 0$ or $p(z) = 0$, (iii) holds trivially. Therefore assume $p(y,z), p(z) > 0$.

We need to bring back probability functions into RHS of (v). Say get $p(y,z)$ involved first, which can be done by integrating over $x$ and substituting (v)

$$p(y,z) = \int p(x,y,z)\,dx = \int f(x,z)g(y,z)\,dx = g(y,z)\int f(x,z)\,dx$$

define $\tilde{f}(z) := \int f(x,z)\,dx$. By $p(y,z) > 0$, either $g(y,z), \tilde{f}(z)$ are both positive or both negative. WLOG assume they are both positive.

$$g(y,z) = \frac{p(y,z)}{\tilde{f}(z)}$$

substituting to (iii) gives

$$p(x,y,z) = \frac{p(y,z)}{\tilde{f}(z)}f(x,z) = \frac{f(x,z)}{\tilde{f}(z)}p(y,z)$$

Luckily, by definition of conditional probability, the fraction is exactly $p(x|y,z)$. Remains to prove $p(x|y,z) = p(x|z)$. Integrate $p(x|y,z)p(y|z) = p(x,y|z)$ w.r.t. $y$

$$\int p(x|y,z)p(y|z)\,dy = p(x|z)$$

$p(x|y,z) = f(x,z)/\tilde{f}(z)$ is independent of $y$, so take it out from the integral,

$$p(x|y,z)\int p(y|z)\,dy = p(x|y,z) = p(x|z)$$

□

When the numerical values of the probabilities are not accessible, or one needs to code such relationships into a computer, logical arguments are required. [16] An abstract concept called *graphoids* is defined for this. It is the logical condition $I(X,Z,Y) \Leftrightarrow X$ and $Y$ are separated by $Z$. In the cased of probability, $I(X,Z,Y) \;\Leftrightarrow\; X \perp Y \,|\, Z$.

**Definition 3** (Graphoid Axioms)**.** Graphoids are logical conditions $I(X,Z,Y)$ that satisfies

- (i) Symmetry: $I(X,Z,Y) \;\Rightarrow\; I(Y,Z,X)$
- (ii) Decomposition: $I(X,Z,Y \wedge W) \;\Rightarrow\; I(X,Z,Y)$ and $I(X,Z,W)$
- (iii) Weak Union: $I(X,Z,Y \wedge W) \;\Rightarrow\; I(X,Z \wedge Y,W)$
- (iv) Contraction: $I(X,Y \wedge Z,W)$ and $I(X,Z,Y) \;\Rightarrow\; I(X,Z,Y \wedge W)$
- (v) Weak closure of intersection: $I(X,Y \wedge Z,W)$ and $I(X,W \wedge Z,Y) \;\Rightarrow\; I(X,Z,Y \wedge W)$

The name "graph"oid comes from the attempts in the last century to represent conditional independence(CI) relations with undirected graphs. It turned out CI are too complicated to be represented by graphs (though it is worth using graphs to provide intuitions)

(ii) means the irrelevance of $X$ to two events jointly can be broken down to the irrelevance of $X$ with each. If $W$ is irrelevant to $X$ conditional on $Z$, (iii) says learning $Y$, which is also irrelevant to $X$ (conditional on $Z$) does not help. $X$ and $W$ are still irrelevant. (conditional on $Z,Y$)

(ii), (iii) and (iv) together convey that $Y,W$ being jointly irrelevant to $X$ is equivalent to the irrelevance of $Y$ and the irrelevance of $W$ after $Y$ is learnt.

(v) says if $W$ is irrelevant to $X$ when you know $Y$ and $Y$ is irrelevant to $X$ when you know $W$, then $Y,W$ are jointly irrelevant to $X$.

Theorem 2.6 in the lecture notes proves CI is a graphoid. Weak closure of intersection is a sweet property only for CI, if the common knowledge $Z$ is dropped, the equation does not hold.

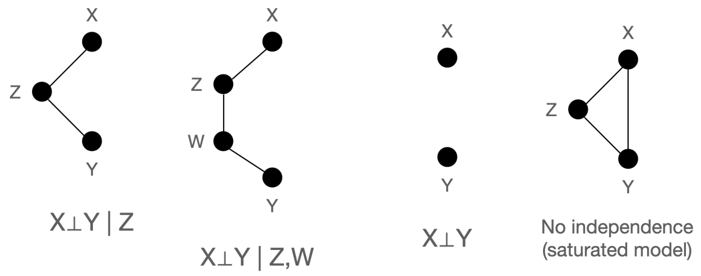


**Figure 1: Representing CI with undirected graphs**

**Exercise 1.** *Check the graphoid axioms make sense by manipulating them on the graphs. e.g. for (i), Z separating X, Y is automatically a symmetric relation.*

Graphical Models are models that utilise conditional independence, so please ensure you understand CI before moving on.

## 1.1 Simple Applications of CI

**Computational Efficiency**
See section 3.5 in the lecture notes.

**Transformation of variable**
Measurable functions are functions whose pre-images are not bizarre, i.e., pre-images are measurable sets in the domain (most sets you can think of are measurable, it is difficult to construct a non-measurable set). e.g. on the real line, all continuous functions are measurable. For $h$ measurable, $Y = y \Rightarrow h(Y) = h(y)$. i.e. learning about $h(Y) = h(y)$ does not change $\{Y = y\}$, $p(x \,|\, h(y), y, z) = p(x \,|\, y, z)$, so

$$X \perp Y \,|\, Z \;\Rightarrow\; X \perp h(Y) \,|\, Y, Z$$

Using theorem 2.6, one can prove learning $h(Y)$ does not change irrelevance of $X, Y$, i.e.

$$X \perp Y \,|\, Z \;\Rightarrow\; X \perp Y \,|\, h(Y), Z$$

and the transformation of a variable by measurable function is possible, i.e.

$$X \perp Y \,|\, Z \;\Rightarrow\; X \perp h(Y) \,|\, Z$$

.

**Sufficient Statistics**
A statistics $T(X)$ for the random variable $X$ from the family $\{P_\theta \,:\, \theta \in \Theta\}$ is sufficient if distribution $X|T(X)$ does not depend on $\theta$. i.e. knowing $T(X)$ allows you to throw away the parameter $\theta$. This sounds like $X \perp \theta \,|\, T(X)$. Indeed, a factorisation theorem states $T(X)$ is sufficient iff

$$f_\theta(x) = g(T(x), \theta)h(x)$$

this corresponds to Theorem 2.4 (v). ($x, \theta$ are packed in different functions through the help of $T(x)$)

## 1.2 Exponential Family

We will follow the form of the exponential family mentioned in the notes. With $x \in \mathscr{X}, \theta \in \Theta$, the pdf for an exponential family is:

$$p(x;\theta) = \exp\left(\sum_i \theta_i \phi_i(x) - A(\theta) - C(x)\right)$$

Statistic $\boldsymbol{\phi}(x) := (\phi_i(x))$ is called sufficient statistic because a theorem states that $\phi(x)$ is always sufficient for an exponential family. The form of the exponential family comes from the attempt to make sufficient statistics $\phi(x)$ to attain the Cramer-Rao lower bound. Note we require the parameter space $\Theta$ to be an open set to allow many useful theorems, and properties to hold. e.g. defining Fisher's information requires differentiation of log-likelihood w.r.t. $\theta$. Non-open sets cause troubles in differentiation.

### 1.2.1 Finding MLE by moment-matching

Lemma 3.1 in the lecture notes states that: surprisingly, the first two derivatives of cumulant function $A(\theta)$ are exactly the mean and covariance of sufficient statistics $\phi(X)$. Following this fact, if one wants to find MLE $\hat{\theta}$ for a member of the exponential family, matching the mean of sufficient statistics $E_{\hat{\theta}}\phi(X)$ with empirical mean $\overline{\phi(X)}$. (the mean of $\phi(X)$ where $X$ are observations) If you have seen moment-matching estimators before, these are straightforward to compute. So calculations are simplified a lot.

### 1.2.2 Multivariate Normal Distribution as Exponential Family

Proving common one-dimensional distributions like Poisson, and geometric distributions are in the exponential family is fairly easy. Adding an exponential and natural logarithm (inverses of each other) is enough. Difficulties might occur regarding multivariate cases: not all expressions can be written in an inner product of vectors.

Think of MVN$(\mu, \Sigma)$ and define the concentration matrix $K := \Sigma^{-1}$. After expanding the exponent, pdf becomes

$$f(x) = C \exp\left\{-\frac{1}{2}x^T Kx + \mu^T Kx + A(K, \mu)\right\}$$

where $A(K, \mu)$ are the terms independent of $x$ and $C$ is some constant independent of $K, \mu, x$. We need to write $K, \mu$ against $x$ in terms of the inner product (of vectors). The term $\mu^T Kx = (K^T\mu)^T x = (K\mu)^T x$ is straight forward. (note: $K$ is symmetric, as covariance matrix $\Sigma$ is symmetric) However, separating $x^T Kx$ in the usual way is difficult. Rather, we use the trick of trace,

$$\begin{aligned} x^T Kx &= \mathrm{Tr}(x^T Kx) \\ &= \mathrm{Tr}(xx^T K) \quad \text{by cyclic property of trace} \\ &= \langle xx^T, K\rangle_F \quad \text{where F means Frobenious norm} \\ &= \mathrm{vec}(xx^T) \cdot \mathrm{vec}(K) \end{aligned}$$

i.e. we treat $xx^T$ and $K$ as vectors (by stacking all rows of the matrix together) and take the inner product of the vectorised matrices. For brevity, we represent $\mathrm{vec}(xx^T), \mathrm{vec}(K)$ by $xx^T$ and $K$. So the parameter $\boldsymbol{\theta} = (\eta := K\mu,\ K)$, the sufficient statistics $\phi(x) = (x,\ -\frac{1}{2}xx^T)$.

Suppose now we need to find the MLE for MVN (assuming $\mu, \Sigma$ are not known), the traditional approach requires differentiating the horrible pdf twice and finding the $\mu, \Sigma$ maximising likelihood. By moment matching, differentiating $A(K, \eta) = \frac{1}{2}\left(\eta^T K^{-1}\eta - \log|K|\right)$ suffices,

$$\nabla_\eta A(K, \eta) = K^{-1}\eta = \mu = E_\theta(\phi_1(x)) = E_\theta(x)$$

$$\nabla_K A(K, \eta) = -\frac{1}{2}K^{-1} - \frac{1}{2}K^{-1}\eta\eta^T K^{-1} = -\frac{1}{2}\Sigma - \frac{1}{2}\mu\mu^T = E_\theta(\phi_2(x)) = E_\theta(-\frac{1}{2}xx^T) = -\frac{1}{2}E_\theta(xx^T)$$

Using empirical means of $E(x) = \overline{x}$ and $E(xx^T) = \overline{xx^T}$, the MLE are $\hat{\mu} = \overline{x}$ and $\hat{\Sigma} = \overline{xx^T} - \hat{\mu}\hat{\mu}^T = \overline{xx^T} - \overline{x}\,\overline{x}^T$

## 1.3 Contingency Tables

Suppose you sampled categorical data from a population, e.g. gender, nationality, religion of students in Oxford. The contingency table summarises the number of samples in each possible combination, e.g. 900 male students from the UK believe in Christianity. All combinations $x_V$ together make a multinomial distribution with likelihood (ignoring normalising constant)

$$\prod_{x_V} p(x_V)^{n(x_V)}$$

where $p(x_V)$ is the probability for combination $x_V$ and $n(x_V)$ is the count. So log-likelihood is

$$l(p) = \sum_{x_V} n(x_V) \log p(x_V)$$

It looks like an exponential family, tempting to define $\theta_{x_V} := \log p(x_V)$. However, in that case, $\Theta = \{\log p(x_V) : x_V \in \mathscr{X}_V\} = (-\infty, 0]$ is not open set as $\log 1 = 0$ is a boundary. So rewrite log-likelihood as

$$l(p) = \sum_{x_V \neq 1_V} n(x_V) \log p(x_V) + \left(1 - \sum_{x_V \neq 1_V} n(x_V)\right) \log p(1_V) = \sum_{x_V \neq 1_V} n(x_V) \log\left(\frac{p(x_V)}{p(1_V)}\right) + \log p(1_V)$$

where $1_V$ is the pattern with 1 in every entry. Now the domain for $\theta_{x_V} := \log p(x_V)/p(1_V)$ $(x_V \neq 1_V)$ is an open set.

### 1.3.1 Log-linear Models

A $n$-way contingency table is a table with $n$ variables. Example 2.5 in the lecture notes is a three-way contingency table. The relations between variables $X, Y$ and $Z$ in a three-way contingency table are worth studying. Let us consider a simple model with $X, Y$ and $Z$ all independent of each other (pairwise independence). Then $\pi_{i,j,k} := P(X = i, Y = j, Z = k) = P(X = i)\,P(Y = j)\,P(Z = k) = \pi_{i,+,+}\pi_{+,j,+}\pi_{+,+,k}$ where $+$ means summing all probabilities over that variable. Since additive models are easier to deal with, we take logarithm:

$$\log(\pi_{i,j,k}) = \log(\pi_{i,+,+}) + \log(\pi_{+,j,+}) + \log(\pi_{+,+,k})$$

Therefore, a feasible model for $\pi_{i,j,k}$ is a log-linear model [4].

$$\log(\pi_{i,j,k}) = \lambda_\emptyset + \lambda_i^A + \lambda_j^B + \lambda_k^C \tag{1}$$

where $\lambda_i^A$ corresponds to $\log(\pi_{i,+,+})$, the other two follow similarly. So what is $\lambda_\emptyset$? Since $\lambda$'s are free parameters rather than the log of probabilities, we need such a constant to ensure $\sum_{i,j,k} \pi_{i,j,k} = N$ ($N$ is the number of observations). $\lambda_\emptyset$ also encodes an overall effect on every cell of the contingency table.

What if now we only have $X \perp Y|Z$? The model (1) does not suffice. The probability formula is

$$P(X = i, Y = j \,|\, Z = k) = P(X = i \,|\, Z = k)P(Y = j \,|\, Z = k)$$

i.e. $\pi_{i,j,k}/\pi_{+,+,k} = (\pi_{i,+,k}/\pi_{+,+,k})(\pi_{+,j,k}/\pi_{+,+,k}) \Rightarrow$,

$$\log(\pi_{i,j,k}) = \log(\pi_{i,+,k}) + \log(\pi_{+,j,k}) - \log(\pi_{+,+,k})$$

so it is worth adding two "interaction terms" to the model (1):

$$\log(\pi_{i,j,k}) = \lambda_\emptyset + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,k}^{X,Z} + \lambda_{j,k}^{Y,Z} \tag{2}$$

$\lambda_{i,j}^{X,Z}$ models the dependence between $X, Z$.

Without any prior information, the model is

$$\log(\pi_{i,j,k}) = \lambda_\emptyset + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,k}^{X,Z} + \lambda_{j,k}^{Y,Z} + \lambda_{i,j}^{X,Y} + \lambda_{i,j,k}^{X,Y,Z} \tag{3}$$

where $\lambda_{i,j,k}^{X,Y,Z}$ encodes the interaction between all three terms. Model (3) contains all possible models. For example, the fourth graph in figure 1 corresponds to $\lambda_{i,j,k}^{X,Y,Z} = 0$.

The model needs to be identifiable, i.e. if two sets of parameters $\lambda$ and $\lambda'$ give the same $\log(\pi_{i,j,k})$, then $\lambda = \lambda'$. Note $\lambda\emptyset$ is free to change, one can always subtract some constant from a $\lambda^{X_V}$ and add that to $\lambda_\emptyset$. So we add a constraint: if any variable equals 1, the corresponding parameter $\lambda^{X_V}$ should be 0 (or any other preferred value).

The model for the two-way contingency table can be found in example 3.5 in the lecture notes. For the $n$-way contingency table with larger $n$, the log-linear model creates too many parameters (the cardinality of the power set is $2^n$). So a different approach is developed. (see [4] Chapter 2)

**Hierarchical Property**: If $\lambda_{i,j,k}^{X,Y,Z} \neq 0$ but $\lambda_{i,j}^{X,Y} = 0$, it would be counter-intuitive. This is saying that $X, Y$ are uncorrelated (conditional on $Z$) but $X, Y, Z$ has some relations. In fact, by theorem 3.7 in the lecture notes, $\lambda_{i,j}^{X,Y} = 0 \Rightarrow \lambda_{i,j,k}^{X,Y,Z} = 0$. So if $\lambda^A = 0$ for some set $A$, any $\lambda^C$ associated with a higher level set $C$ ($A \subseteq C$) is 0. So, the log-linear model for the contingency table is *hierarchical*.

**Top-down Approach**: When fitting log-linear in practice, a saturated model is fitted at first. Then, a lower-level model with fewer parameters can be fitted for comparison. e.g. reducing model (3) to model (2). Deviance [4] (basically log-likelihood ratio test) is used to check the goodness of fit of the reduced model compared to the full model:

$$G^2 = 2(\hat{l}_f - \hat{l}_r)$$

where $\hat{l}_f$ is log-likelihood of full model and $\hat{l}_r$ is the log-likelihood of the reduced model. Some conditional independence can be inferred if the reduced model is better.

## 2 Undirected Graphs

When introducing graphoid (definition 3), the intuition "$X, Y$ are separated by $Z$" is used for $X \perp Y \mid Z$. This can be realised using a graph. This chapter focuses on how CI can be fitted onto graphs.

Only simple, undirected graphs are used in this section, please see 4.1 in the lecture notes for the basic definitions. The formalisation of "$X, Y$ are separated by $Z$" is

**Definition 4** (Separation)**.** sets of vertices $A, B$ are separated by $S \subseteq V$ if every path from $a \in A$ to $b \in B$ contains at least one vertex in $S$. This is denoted $A \perp_s B \mid S$ in $\mathcal{G}$.

Given a group of random variables $\{X_i\}_{i=1,\cdots,n}$ with distribution $p$. A graph with $|V|= n$ nodes is required, and let $X_i$ correspond to node $i \in V$. There are three ways to define the idea: the graph $\mathcal{G}$ correctly describes CI between the variables using "separation". The easiest one is *pairwise Markov property.*

**Definition 5** (pairwise Markov property)**.** Distribution $p$ satisfies *pairwise Markov property* for undirected graph $\mathcal{G}$ if

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

Nodes $i, j$ not connected implies $X_i, X_j$ being independent conditional on the rest of variables.

This definition can be extended to groups of nodes, using definition4.

**Definition 6** (global Markov property)**.** Distribution $p$ satisfies *global Markov property* for undirected graph $\mathcal{G}$ if for any disjoint sets of nodes $A, B, S \subseteq V$,

$$A \perp_s B \mid S \text{ in } \mathcal{G} \Rightarrow X_A \perp X_B \mid X_S[p]$$

By selecting $A = \{i\}$, $B = \{j\}$ and $S = V \setminus \{i, j\}$, one can see that the global Markov property implies the pairwise Markov property.

Note if $A, B, S$ are not disjoint, the case becomes trivial. e.g. select $A = 1$, $B = 2$, $S = 1$, then $X_1 \perp X_2 \mid X_1$ is trivially true. So, usually, we are interested in the case of $A, B, S$ being disjoint.

A complete set of nodes (each pair of nodes is joined by an edge) means no independence between variables. But in this sense, CI only exists between the complete parts of the graph.

**Definition 7** (clique)**.** A clique is a maximal complete set, where maximal means that adding any other node breaks completeness. The set of all cliques in $\mathcal{G}$ is denoted as $\mathcal{C}(\mathcal{G})$.

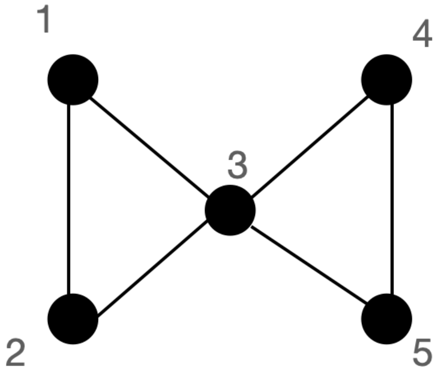


**Figure 2: Example Graph with two cliques**

In figure 2, $\{1,2,3\}$ and $\{3,4,5\}$ are the two cliques. So cliques are not necessarily disjoint. The joining point can be viewed as the condition event, which separates all other events in the two cliques.

The third definition for the relation between graph $\mathcal{G}$ and CI is a factorisation over the set of cliques.

**Definition 8** (Factorisation on graph)**.** *$p$ factorises according to $\mathcal{G}$* if

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

for some functions $\phi_C$ (called *potentials*)

This definition sounds very strong as it puts functions that behave like probabilities onto the graph. Indeed, factorisation implies global Markov property.

**Theorem 2.1.** *If $p(x_V)$ factorises according to $\mathcal{G}$, then $p$ satisfies global Markov property for undirected graph $\mathcal{G}$.*

The converse does not hold. Problem sheet B3 is a counter-example.

*Proof.* This concise proof is given in [10].

Suppose $S$ separates $A, B$ ($S, A, B$ are disjoint). The goal is to use the factorisation criterion to prove $X_A \perp X_B | X_S$, i.e. to separate the probabilities involving $A, B$. We already have a factorisation over cliques. Let $\mathcal{C}(\mathcal{G})_A$ be the set of cliques in $\mathcal{C}(\mathcal{G})$ intersecting with the set $A$. For $C \in \mathcal{C}(\mathcal{G})_A$, $B \cap C = \emptyset$, as otherwise, we have an edge joining $A, B$. (note $C$ is complete) Similarly, for $C \in \mathcal{C}(\mathcal{G}) \setminus \mathcal{C}(\mathcal{G})_A$, $A \cap C = \emptyset$
So the factorisation becomes

$$p(x_V) = \prod_{C \in \mathcal{C}(\mathcal{G})_A} \phi(x_C) \prod_{C \in \mathcal{C}(\mathcal{G}) \setminus \mathcal{C}(\mathcal{G})_A} \phi(x_C) = f(x_{S \cup A}) g(x_{S \cup B})$$

□

We already proved (factorisation ⇒ global Markov ⇒ pairwise Markov). One theorem finishes the loop.

**Theorem 2.2** (Hammersley-Clifford Theorem)**.** *If $p(x_V) > 0$ satisfies pairwise Markov property for $\mathcal{G}$, then $p$ factorises according to $\mathcal{G}$. (see [9], Theorem 3.9)*

So, if the probability is strictly positive, three definitions are equivalent. The fact that (pairwise Markov ⇒ global Markov) is a bit like hierarchical log-linear models. A lower-order independence ($\lambda_A = 0$) implies the higher-order independence. ($\lambda_B = 0$ where $A \subseteq B$) Still, the theorem is quite remarkable.

## 2.1 Decomposability

We have seen in the last section that cliques and complete sets are important for separating the graph and Markov properties. In this section, graph decomposability will be studied.

Decomposable graphs are the ones which can be broken down into several complete parts that are possibly linked with each other. This can be achieved by a series of binary "decompose" operations.

**Definition 9** (decomposition)**.** If $A \cup S \cup B = V$ and $A, B, S$ are disjoint, $(A, S, B)$ is a decomposition if

- $A \perp_S B$, i.e. $A, B$ are separated by $S$
- $\mathcal{G}_S$(sub-graph generated by $S$) is complete.

Look at graph $\mathcal{G}$ in figure 3, $(\{1,2,3\}, \{4\}, \{5,6\})$ is a decomposition. The graph is decomposed into $\mathcal{G}_1 := \mathcal{G}_{\{1,2,3,4\}}$, $\mathcal{G}_2 := \mathcal{G}_{\{4,5,6\}}$. $\mathcal{G}_2$ is already complete, but $\mathcal{G}_1$ can be further decomposed. This is done using decomposition $(\{1\}, \{2,3\}, \{4\})$.

It is time to define *decomposable graph* rigorously.

**Definition 10** (decomposable)**.** A graph $\mathcal{G}$ is decomposable either if one of the following holds

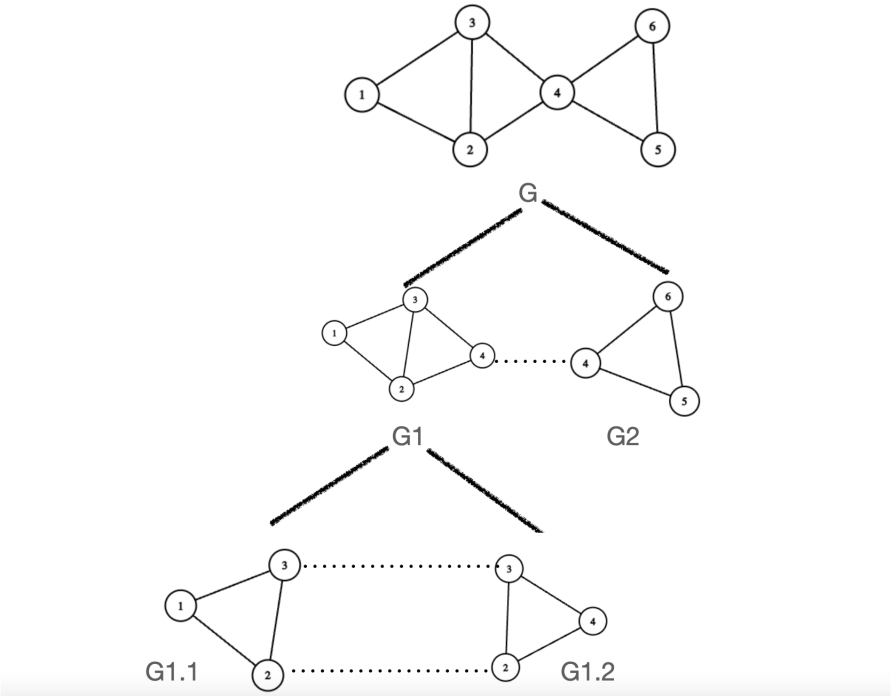


**Figure 3: A decomposable graph.**
**Solid line: decomposition process. Dotted line: links nodes that were the same in the original graph.**

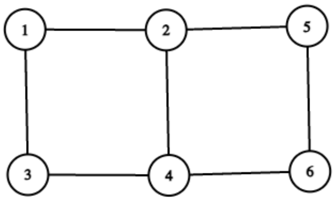


**Figure 4: A non-decomposable graph**

- $\mathcal{G}$ is complete
- $\mathcal{G}$ has decomposition $(A, S, B)$ and $\mathcal{G}_{A\cup S}$, $\mathcal{G}_{B\cup S}$ are decomposable.

Note this is an iterative definition. Hopefully, the idea is clear from figure 3. Note that having a decomposition is NOT decomposable. See the following example.

**Example 1.** The graph in figure 4 has a decomposition $(\{1,3\},\{2,4\},\{5,6\})$. However, the subgraph $\mathcal{G}_{1,2,3,4}$ is not decomposable. This graph corresponds to B3 in problem sheet 1, where we have proved the probability cannot be decomposed upon the cliques in part b.

By comparing figure 3 and 4, what is the main difference? I will leave the answer after introducing the running intersection property.

Consider a general clustering of the nodes, $\mathcal{C} := \{C_1, \cdots, C_k\}$ where $C_i \subseteq V$. (Note: the set of cliques is one special way of clustering nodes) We can build the *cluster graph* based on $\mathcal{C}$. Roughly speaking, it is a graph with $C_i$ being the nodes, and $C_i \cap C_j$ being edge $\{i, j\}$ if the intersection is not empty.

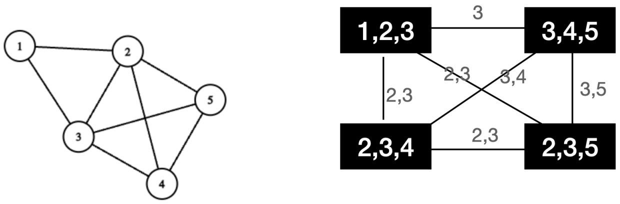


**Figure 5: Cluster graph of a graph**

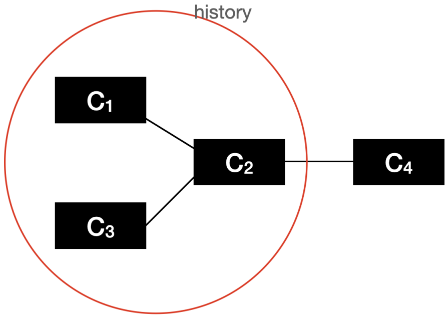


**Figure 6: Cluster graph of a graph satisfying RIP**

**Example 2.** For the graph to the left of figure 5, consider the clusters $\{1, 2, 3\}, \{2, 3, 4\}, \{2, 3, 5\}, \{3, 4, 5\}$. The corresponding clustering tree is shown on the right.

The running intersection property (RIP) says there is an ordering $C_1, \cdots, C_k$ of $\mathcal{C}$ such that the set $S_j := C_j \cap H_{j-1}$ where the history set $H_{j-1} := \cup_{i=1}^{j-1} C_i$ (elements of $C_j$ that are already covered by history sets) is completely contained in one set in the history $C_{\sigma(j)}$ $(\sigma(j) < j)$. i.e. $S_j = C_j \cap C_{\sigma(j)}$. Note $S_j$ is also called the separator, and $R_j = S_j \setminus H_{j-1}$ is called the residuals. By definition, $R_j \cup S_j = C_j$.

**Example 3.** Figure 6 demonstrates the graphical meaning of RIP. When $j = 4$, the entire history $H$ is in the red circle. History of $C_4$, i.e. $C_4 \cap H$, comes from $C_3$ ONLY. Therefore, $C_4 \cap H = C_4 \cap C_3$. Similarly, when $j = 3$, history of $C_3$ all comes from $C_2$. However, note that if you change the order of $C_2, C_3$, the property does not hold (why?). RIP only requires one ordering where this property holds, so the graph satisfies RIP.

**Example 4.** Let's consider the notorious graph to the left of figure 7 with clustering $\{1, 2\}, \{2, 4\}, \{1, 3\}, \{3, 4\}$. No matter how you give orders to them, you will face the situation in the right of figure 7 when $j = 4$. The history of $C_4$ comes from two sets in the red circle. So, the graph does not satisfy RIP.

Now, the main theorem of this section can be presented:

**Theorem 2.3.** *For undirected graph $\mathcal{G}$, the following are equivalent*

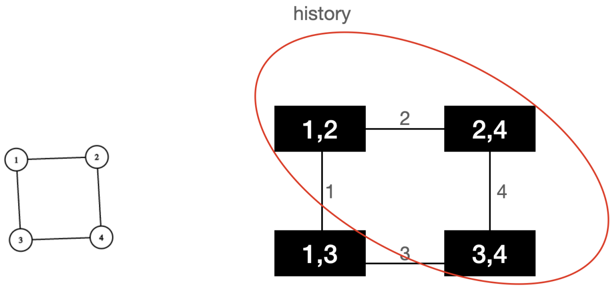


**Figure 7: Clustering tree of a graph NOT satisfying RIP**

- *(i) $\mathcal{G}$ is decomposable*
- *(ii) $\mathcal{G}$ is triangulated (there are triangles everywhere on the graph, but no squares, no pentagon, no $n$-gon with $n > 3$)*
- *(iii) For all $a, b \in V$, every minimal $a, b$-separator (i.e. the set $S$ that separates $\{a\}, \{b\}$, but any proper subset of $S$ cannot separate $\{a\}, \{b\}$) is complete*
- *(iv) $\mathcal{C}(\mathcal{G})$ satisfies RIP*

(ii) is the main difference between the graph in figure 3 and 4 I mentioned before. (iii) makes the definition of decomposable local, i.e. you only need to prove separation exists for singletons $A = \{a\}, B = \{b\}$. (iv) is about the ordering of cliques, which has something to do with the ordering of decomposition as in figure 3.

*Proof.* **(i) ⇒ (ii)**
$\mathcal{G}$ being decomposable has a trivial case: $\mathcal{G}$ is complete. Complete graphs are triangulated. Otherwise, there is a decomposition into $\mathcal{G}_{A\cup S}$, $\mathcal{G}_{B\cup S}$ possible.
Decomposition naturally reduces the number of vertices (in each sub-graph), so we do induction on the number of vertices $p$.
Base case: Any graph with $p < 4$ is triangulated by definition.
Induction: by the inductive hypothesis, $\mathcal{G}_{A\cup S}$, $\mathcal{G}_{B\cup S}$ are both triangulated, so any circle with length $\geq 4$ must involve both $A$ and $B$, say $a \in A, b \in B$ are on the circle. $S$ separates $A$ and $B$, so this circle must at least involve two points $s_1, s_2 \in S$ on the paths $a \to b$ and $b \to a$ respectively. But $S$ is complete by the definition of decomposition, so $s_1, s_2$ are connected. But this is a chord on the circle, a contradiction.

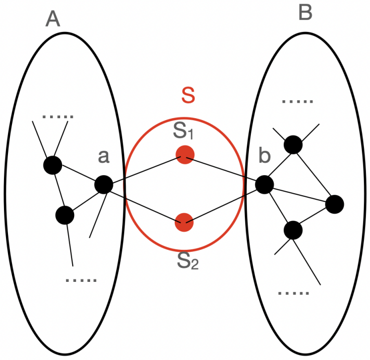


**Figure 8: Graphical illustration for proving(i) ⇒ (ii)**

**(ii) ⇒ (iii)**

For the sake of contradiction, assume there is a minimal $a, b$-separator, but $s_1, s_2$ are not adjacent. By minimality, there is at least one path from $a$ to $b$ passing $s_1$ but not $S \setminus \{s_1\}$, and another path passing $s_2$ but not $S \setminus \{s_2\}$. This makes a circle $(a, \cdots, s_1, \cdots, b, \cdots, s_2, \cdots, a)$. There may be chords on the circle, but no chord on $s_1, s_2$.

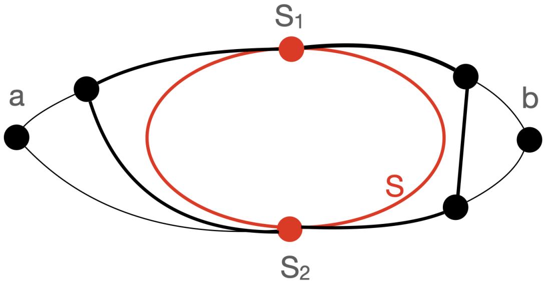


**Figure 9: Graphical illustration for proving(ii) ⇒ (iii). The minimal traversed circle is in bold**

One can traverse all the chords to shorten the circle, the circle with minimal length has length $> 4$ due to the separation of $S$. (as shown in figure 9).

**(iii) ⇒ (iv)**

Suppose the graph is complete, then there is only one clique, naturally satisfying RIP. Otherwise, there exist nodes $a, b$ which are not adjacent. If $a, b$ has no separator, then the graph is at least bipartite (there are at least two non-connecting parts) We can find the RIP of cliques on each connected component separately, and concatenate the sequences, which still satisfies RIP (why?)

Otherwise, if $a, b$ has a separator, reduce it to a minimal separator $S$. By (iii), $S$ is complete, so $S \subseteq C_i$ for some $i$. The rest of the proof is part of problem sheet 2.

**(iv) ⇒ (i)**

Do induction on the number of cliques.
Base case: one clique (graph is complete), nothing to prove
Inductive step: suppose $C_1, \cdots, C_k$ is an ordering of all the cliques satisfying RIP. (WLOG assume that $C_i$ are distinct)
**Claim**: $S_k := C_k \cap H_{k-1}$ separates $R_k := C_k \setminus H_{k-1}$ and $H_{k-1} \setminus S_k$.
Note that $S_k \cup R_k \cup (H_{k-1} \setminus S_k) = H_k = V$ as all the cliques comprise the whole graph.

Suppose for contradiction that there is $a \in R_k$ and $b \in H_{k-1} \setminus S_k$ such that $a, b$ are adjacent (see figure 10). $\{a, b\}$ is a complete set, so $\{a, b\} \subseteq C_p$ for some clique $C_p$. But $b \notin C_k$, so $p \neq k$. $a \notin H_{k-1}$, so $p \neq 1, \cdots, k-1$. Contradiction! Such an edge cannot exist. So separation holds.
Also note that by RIP, $S_k \subseteq C_j$ for some $j > k$, if $R_k = \emptyset$, then $C_j = C_k$, but $C_i$ are distinct. So $R_k \neq \emptyset$, which means $\mathcal{G}_{H_{k-1}}$ only has $k-1$ cliques. By the inductive hypothesis, $\mathcal{G}_{H_{k-1}}$ is decomposable, also, $\mathcal{G}_{R_k \cup S_k} = \mathcal{G}_{C_k}$ is complete, so decomposable as well.

□

In the last part of the proof of Theorem 2.3, we showed that $S_k$ separates the remainder set and the set in history other than $S_k$. Hence, $S_k$ is called the $k$th separator. If $C_1, \cdots, C_k$ is an ordering of cliques satisfying RIP, then $S_2, \cdots, S_n$ are separators. By default, $S_1 := \emptyset$.

Separators help to break down probabilities, as the following lemma suggests

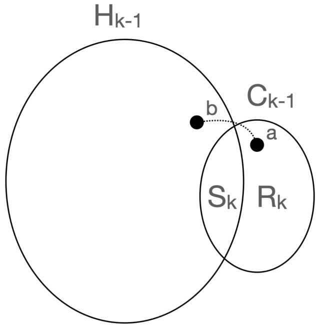


**Figure 10: Graphical illustration for proving(iv) ⇒ (i), the dotted line is a forbidden edge**

**Lemma 2.4.** *Given decomposition $(A, S, B)$ for graph $\mathcal{G}$ and distribution $p$,*

$$p \textit{ factorises w.r.t } \mathcal{G} \;\Leftrightarrow\; p(x_V)p(x_S) = p(x_{A\cup S})p(x_{B\cup S})$$

Applying the lemma to $k$th separator, if $p$ factorises w.r.t. $\mathcal{G}$,

$$p(x_{S_k})p(x_V) = p(x_{C_k})p(x_{H_{k-1}}) \;\Rightarrow\; p(x_V) = \frac{p(x_{C_k})}{p(x_{S_k})}p(x_{H_{k-1}})$$

continue factorising $H_{k-1}$, and repeat until $H_1$, we obtain a factorisation formulae over separators,

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

**Proposition 2.5.** *$p$ factorises w.r.t. decomposable graph $\mathcal{G}$ iff*

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

*Further, $C_i \setminus S_i$ are disjoint, so $p(x_{C_i\setminus S_i} \,|\, x_{S_i})$ can jointly take any set of values, there is no constraint.*

Applying this to the log-linear model,

$$\begin{aligned} l(p) &= \sum_{x_V} n(x_V) \log \left( \prod_{i=1}^{k} p(x_{C_i\setminus S_i} \,|\, x_{S_i}) \right) \\ &= \sum_{i=1}^{k} \sum_{x_V \in \mathscr{X}_V} n(x_V) \log p(x_{C_i\setminus S_i} \,|\, x_{S_i}) \\ &= \sum_{i=1}^{k} \sum_{x_{C_i} \in \mathscr{X}_{C_i}} n(x_{C_i}) \log p(x_{C_i\setminus S_i} \,|\, x_{S_i}) \end{aligned}$$

the last equality holds because the term $\log p(x_{C_i\setminus S_i} \,|\, x_{S_i})$ remains the same as long as entries of $x_V$ for variables in $C_i$ are the same, that means

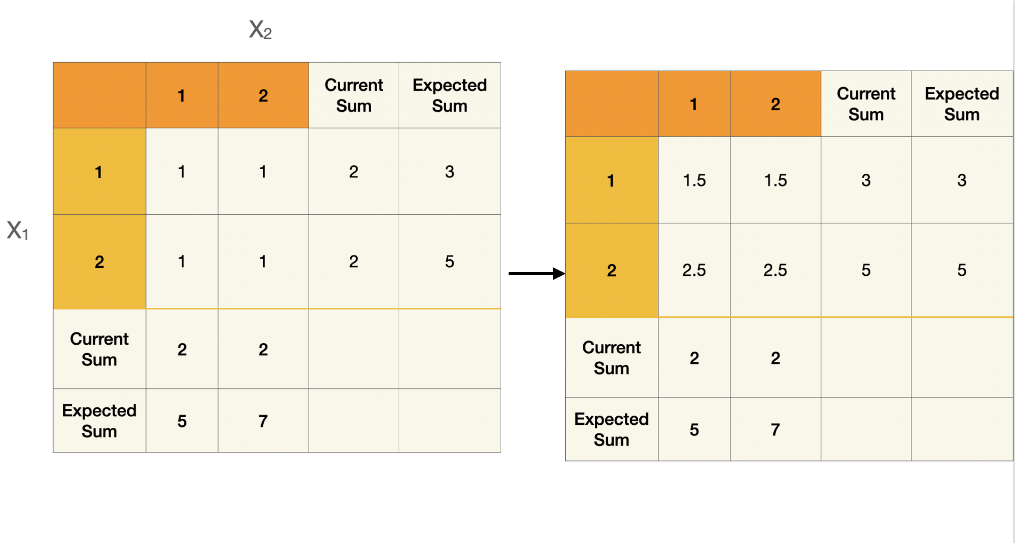
$$\sum_{x_{V\setminus C_i} \in \mathscr{X}_{V\setminus C_i}} n(x_V) \log p(x_{C_i\setminus S_i} \,|\, x_{S_i}) = n(x_{C_i}) \log p(x_{C_i\setminus S_i} \,|\, x_{S_i})$$

so

$$\sum_{x_V \in \mathscr{X}_V} n(x_V) \log p(x_{C_i\setminus S_i} \,|\, x_{S_i}) = \sum_{x_{C_i} \in \mathscr{X}_{C_i}} \sum_{x_{V\setminus C_i} \in \mathscr{X}_{V\setminus C_i}} n(x_V) \log p(x_{C_i\setminus S_i} \,|\, x_{S_i}) = \sum_{x_{C_i} \in \mathscr{X}_{C_i}} n(x_{C_i}) \log p(x_{C_i\setminus S_i} \,|\, x_{S_i})$$

Convince yourself using a simple graph $V = \{0, 1\}$ with no edge (two independent variables, so the contingency table has 4 cells), checking with the cliques $\{0\}$ and $\{1\}$.

The final line in the equation means that all inferences can be made on the margins $C_i$. However, there is no such formulae for a non-decomposable model. In such cases, IPF is used to match the probabilities with given marginal probabilities $q(x_{C_i})$. Usually, we start with the discrete uniform distribution for cells on the contingency table, as all variables are independent in this case. So Markov property is trivially satisfied.



| | 1 | 2 | Current Sum | Expected Sum |
|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 3 |
| 2 | 1 | 1 | 2 | 5 |
| Current Sum | 2 | 2 | | |
| Expected Sum | 5 | 7 | | |

| | 1 | 2 | Current Sum | Expected Sum |
|---|---|---|---|---|
| 1 | 1.5 | 1.5 | 3 | 3 |
| 2 | 2.5 | 2.5 | 5 | 5 |
| Current Sum | 2 | 2 | | |
| Expected Sum | 5 | 7 | | |

**Figure 11: Iterative Proportional Fitting for two-way contingency table**

The idea is simple, multiply values in each margin by a proportion constant to correct the marginal sum. For example, in figure 11, adjust $X_1$ first, the row sum of $X_1 = 1$ is 2, whereas the target is 3, so the row should be multiplied with $3/2 = 1.5$. After fitting the rows, deal with $X_2$ (columns). This will alter the row sums obtained before, but we are getting closer to the final answer. So fit the rows again, then the columns again. Looping this over margins iteratively ensures $p(x_C)$ converges to $\hat{p}(x_C) := n(x_C)/n$, the MLE estimator among all distributions Markov w.r.t. $\mathcal{G}$. The example given in the lecture notes is a four-way contingency table, hope the process makes sense after looking at the two-way table in figure 11.

# 3 Directed Graphs

Directed graphs are common in casual inference. And edge $a \to b$ means $a$ causes $b$. Similarly to undirected graphs, we can study how to fit probabilities to directed graphs. All graphs in this section are assumed to be DAG.

Instead of studying cliques, factorisation on DAG conditions on parents, i.e.

$$p(x_V) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)})$$

This implies the *local Markov property*: $X_v$ is independent of $v$'s non-descendants except parents of $v$, conditional on the parents.

A straightforward result is that factorisation on $\mathcal{G}$ implies factorisation on $\mathcal{G}_A$ as long as $A$ contains all of its own parents. (i.e. $A$ contains all ancestors, or $A$ is *ancestral*)

**Proposition 3.1.** *If $p$ factorises over $\mathcal{G}$, and $A$ is ancestral set, then $p$ also factorises over $\mathcal{G}_A$.*

But factorisation for undirected graphs cannot be passed to the subgraph, as removing nodes changes the set of cliques $\mathcal{C}(\mathcal{G})$.

There are three types of local structure for DAG [28], see figure 12.

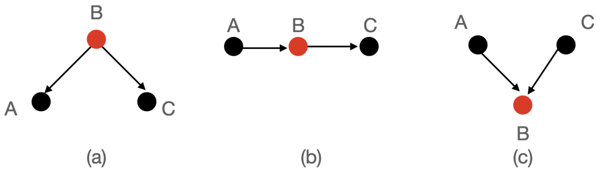


**Figure 12: Three local structures of DAG**

(a) translates into $B$ causes $A, C$, two independent factors. Applying local Markov property on node $A$ implies $A \perp C \mid B$. (b) is a chain of causalities, applying local Markov property to node $C$ yields $C \perp A \mid B$. (c) looks similar to the first two, but $A, C$ are not independent unless $B$ is not present. Because $A$ and $C$ causes $B$ together. e.g. congestion $A$ and weather on the day $C$ are independent, but they affect the bus arrival time $B$ together. If $B$ is given, $A$ and $C$ become related through their common descendant $B$. For example, a bus arriving late on sunny weather could mean a higher probability of congestion on the roads.

Therefore, when restoring the moral graph $\mathcal{G}^m$ (changing a directed graph to an undirected graph), if there is a structure like (c), called *v-structure*, an additional edge will be given to $A$ and $C$ as they are not conditionally independent. For structures like (a) and (b), the moral graph removes the direction of edges.

**Proposition 3.2.** *Factorisation of $p$ over DAG $\mathcal{G}$ implies factorisation over $\mathcal{G}^m$*

*Proof.* By definition

$$p(x_V) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)}) = \prod_{v \in V} \phi(x_{v \cup \mathrm{pa}(v)})$$

where $\phi(x_{v \cup \mathrm{pa}(v)}) := p(x_v \mid x_{\mathrm{pa}(v)})$. Note the moral graph on $v \cup \mathrm{pa}(v)$ is complete, as any pair of disjoint parents of $v$ makes a v-structure, so they are all joined in the moral graph. Therefore, $\mathcal{G}^m_{v \cup \mathrm{pa}(v)}$ are complete sets in $\mathcal{G}$. Complete sets are contained in cliques, so the above expression is also a factorisation over the cliques in $\mathcal{G}^m$. □

The *global Markov property* for DAG is defined on the moral graph

**Definition 11** (global Markov property)**.** $p$ satisfies global Markov property w.r.t to $\mathcal{G}$ if

$$A \perp_M C \,|\, B \;\Rightarrow\; X_A \perp X_B \,|\, X_C[p]$$

where $M := (\mathcal{G}_{\mathrm{an}(A\cup B\cup C)})^m$.
So it is equivalent to global Markov property for the undirected moral graph of the ancestral set.
The whole moral graph $\mathcal{G}^m$ is not used. By Proposition 3.1, studying the union of sets and all its ancestors is enough for DAG.

Note global Markov property means (separation on the moral graph ⇒ independence w.r.t. $p$). But is there any independence not revealed on the graph by separation?

The answer is yes. However, if $X_A \perp X_B \,|\, X_C$, and $A$ and $B$ are not separated by $C$ in the moral graph $M$, then $X_A \perp X_B \,|\, X_C[p]$ is not true for all $p$ satisfying the global Markov property. Taking the contra-positive argument: exists $p$ satisfying global Markov property such that (independence w.r.t. $p \Rightarrow$ separation on the moral graph), i.e. exists $p$ s.t.

$$A \perp_M C \,|\, B \;\Leftrightarrow\; X_A \perp X_B \,|\, X_C[p]$$

for all sets $A, B, C$. This means for that $p$, all independence can be found by separations on the moral graph. It is called the *completeness* of global Markov property.

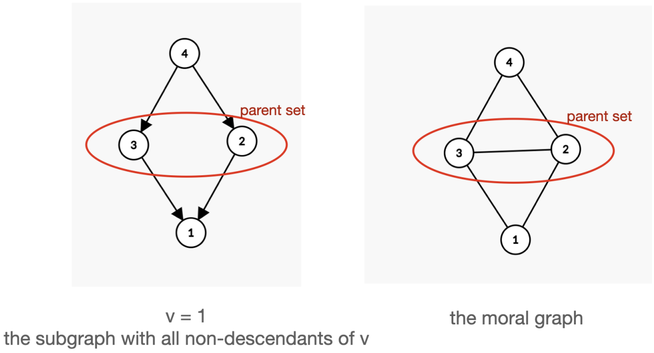
With the three properties: factorisation, global Markov, and local Markov, it is worth studying their relationships. Surprisingly, they are all equivalent even without the condition of $p$ being strictly positive. (This condition is required for factorisation, global Markov and pairwise Markov properties to be equivalent on undirected graphs.)

**Theorem 3.3.** *Given DAG $\mathcal{G}$ and probability $p$, the following are equivalent:*

- *(i) $p$ factorises according to $\mathcal{G}$*
- *(ii) $p$ is global Markov w.r.t. $\mathcal{G}$*
- *(iii) $p$ is local Markov w.r.t. $\mathcal{G}$*

*Proof.* **(i) ⇒ (ii)**
The aim is to use the global Markov property on the moral graph $(\mathcal{G}_W)^m$, where $W = \mathrm{an}_{\mathcal{G}}(A\cup B\cup C)$. Recall factorisation can be passed to the moral graph and subgraph by proposition 3.2 and 3.1. So $p$ factorises over $(\mathcal{G}_W)^m$, which means it satisfies the global Markov property w.r.t. $(\mathcal{G}_W)^m$.

**(ii) ⇒ (iii)**
By definition, it is enough to argue $v$ and $\mathrm{nd}_{\mathcal{G}}(v)\backslash\mathrm{pa}_{\mathcal{G}}(v)$ are separated by $\mathrm{pa}_{\mathcal{G}}(v)$ on the moral graph $(\mathcal{G}_{\{v\}\cup\mathrm{nd}_{\mathcal{G}}(v)})^m$.



v = 1
the subgraph with all non-descendants of v

the moral graph

**Figure 13: Illustration of (ii) implies (iii). The left is a subgraph induced by the non-descendants of $v = 1$: $\mathcal{G}_{\mathbf{nd}(1)\cup\{1\}}$, and the right is its moral graph.**

First, there could not be a path from $\mathrm{nd}_{\mathcal{G}}(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)$ to $v$ without involving $\mathrm{pa}_{\mathcal{G}}(v)$ on $\mathcal{G}_{\{v\}\cup\mathrm{nd}_{\mathcal{G}}(v)}$ by definition of

parent set $\mathrm{pa}_{\mathcal{G}}(v)$. (on figure 13, node 4 cannot go to node 1 without passing parents of 1)
Any additional edge on the moral graph is added to parents sharing the same descendant, but $v$ has no descendant in $\{v\} \cup \mathrm{nd}_{\mathcal{G}}(v)$. So $v$ is still separated from $\mathrm{nd}_{\mathcal{G}}(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)$ on the moral graph.

**(iii) ⇒ (i)**
Define previous set $\mathrm{pre}_<(v)$ for an ordering $<$ to be $\{w \mid w < v\}$. By law of total probability,

$$p(x_v)) = \prod_v p(x_v \mid x_{\mathrm{pre}_<(v)})$$

looking back at the definition of local Markov property: $X_v$ is independent $X_{\mathrm{nd}_{\mathcal{G}}(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)}$ given $X_{\mathrm{pa}_{\mathcal{G}}(v)}$. If we assume $<$ is topological order, then any parent of $v$ comes before $v$, and any descendant of $v$ is placed after $v$. So $\mathrm{pre}_<(v) \setminus \mathrm{pa}_{\mathcal{G}}(v) \subseteq \mathrm{nd}_{\mathcal{G}}(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)$. By decomposition axiom, $X_v$ is independent of $X_{\mathrm{pre}_<(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)}$ given $X_{\mathrm{pa}_{\mathcal{G}}(v)}$, which means $p(x_v \mid x_{\mathrm{pre}_<(v)}) = p(x_v \mid x_{\mathrm{pa}(v)})$. Substituting back to the sum yields

$$p(x_v)) = \prod_v p(x_v \mid x_{\mathrm{pa}(v)})$$

□

By the theorem above, whenever we say $p$ is Markov w.r.t. $\mathcal{G}$, it could mean $p$ factorises according to $\mathcal{G}$ or the global Markov property or the local Markov property.

**Contingency table and DAG**

Similar to undirected graphs, the factorisation formulae can be used to simplify the log-linear model. Suppose $p$ factorises over DAG $\mathcal{G}$,

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

once again, note $\log p(x_v \mid x_{\mathrm{pa(v)}})$ is constant over $V \setminus (\{v\} \cup \mathrm{pa(v)})$, therefore we can rearrange the summation

$$\begin{aligned} &= \sum_{v \in V} \sum_{x_{v \cup \mathrm{pa(v)}}} \log p(x_v \mid x_{\mathrm{pa(v)}}) \sum_{x_{V \setminus (\{v\} \cup \mathrm{pa(v)})}} n(x_V) \\ &= \sum_{v \in V} \sum_{x_{v \cup \mathrm{pa(v)}}} \log p(x_v \mid x_{\mathrm{pa(v)}}) n(x_{v \cup \mathrm{pa(v)}}) \end{aligned}$$

So likelihood on $\{v\} \cup \mathrm{pa(v)}$ can be studied individually for each $v \in V$. And by the fact that MLE of the contingency table is the empirical count ratio,

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

so the joint distribution

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

**Bayes model on contingency table**:
Suppose the prior $\pi(\theta) = \prod_v \pi(\theta_v)$ (independent prior for each variable), then

$$\begin{aligned} \pi(\theta \mid x_v) &\propto \pi(\theta) p(x_V \mid \theta) \\ &= \prod_v \pi(\theta_v) p(x_v \mid x_{\mathrm{pa}(v)}, \theta_v) \end{aligned}$$

so each term is independent, and $\pi(\theta_v \mid x_V) = \pi(\theta_v \mid x_v, x_{\mathrm{pa}(v)})$. So update of $\theta_v$ is based purely on the set $\{v\} \cup \mathrm{pa}(v)$.

## 3.1 Markov Equivalence

Suppose $\mathcal{G}$ and $\mathcal{G}'$ are two different undirected graphs defined on the set of nodes $V$. And edge $(i, j)$ is present in $\mathcal{G}'$ but not in $\mathcal{G}$. Then any $p$ that is Markov on the $\mathcal{G}$ satisfies $X_i \perp X_j \mid X_{V\setminus\{i,j\}}$, but $p$ is not Markov on $\mathcal{G}'$ as $i, j$ are adjacent. So different undirected graphs have different Markov probability models.

However, graphs (a) and (b) of figure 12 imply the same independence: $A \perp C \mid B$. Such graphs are called Markov equivalent.

**Definition 12** (Markov Equivalent)**.** $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if any $p$ Markov w.r.t. $\mathcal{G}$ is also Markov w.r.t $\mathcal{G}'$

If we treat all directed edges as undirected, (a) and (b) are essentially the same graph. But (c) also becomes the same graph if we remove the direction of edges, while $A \not\perp C \mid B$ on (c). In fact, (c) is single, meaning no other directed graph on $\{A, B, C\}$ is Markov equivalent to (c). Equivalently, the v-structure is distinguishable from the other structures.

**Definition 13** (Skeleton)**.** For DAG $\mathcal{G}$, the skeleton $\mathrm{skel}(\mathcal{G})$ is an undirected graph formed by ignoring the directions of edges in $\mathcal{G}$.

Note the only difference between $(\mathcal{G})^m$ and $\mathrm{skel}(\mathcal{G})$ is: $(\mathcal{G})^m$ adds additional edges to parents sharing the same descendant node.

As discussed above, having the same skeleton does not imply Markov equivalence, but the converse is true.

**Lemma 3.4.** *Markov equivalent DAGs must have the same skeleton.*

*Proof.* Proving two graphs that are not Markov equivalent is easier. Prove the contra-positive statement instead. If $\mathrm{skel}(\mathcal{G}) \neq \mathrm{skel}(\mathcal{G}')$, we aim to construct a $p$ that is Markov on $\mathcal{G}$ but not $\mathcal{G}'$. From which $\mathcal{G}$ and $\mathcal{G}'$ are not Markov equivalent follows.

Suppose $i \to j$ on $\mathcal{G}$ but $i$ and $j$ are not adjacent on $\mathcal{G}'$. One cheating technique to define $p$ satisfying Markov on $\mathcal{G}$ is to force everything independent except $X_i$ and $X_j$ so that we can focus on $\{i, j\}$. Because Markov property says (separation ⇒ independence), but NOT (independence ⇒ separation). Adding redundant independence does not break Markov's property. Clearly, $p$ with

- $X_i \not\perp X_j [p]$
- $X_v \perp X_{V\setminus\{v\}} [p]$ for all $v \neq i, j$

is Markov w.r.t. $\mathcal{G}$. Since $i$ is $j$'s parent, there is no independence requirement between $i$ and $j$.

To construct such $p$, by the law of total probability, as long as $p$ satisfies the following factorisation, the assumptions are satisfied:

$$p(x_V) = p(x_j \mid x_i) \prod_{v \in V\setminus\{j\}} p(x_v)$$

where $p(x_j \mid x_i) \neq p(x_j)$ for some $x_j$.

However, such $p$ is not Markov w.r.t. $\mathcal{G}'$. Suppose for contradiction it is Markov w.r.t. $\mathcal{G}'$, by local Markov property,

$$X_i \perp X_{\mathrm{nd}(i)\setminus\mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)} \quad \text{and } X_j \perp X_{\mathrm{nd}(j)\setminus\mathrm{pa}(j)} \mid X_{\mathrm{pa}(j)}$$

$\mathcal{G}'$ is acyclic so either $i \in \mathrm{nd}(j)$ or $j \in \mathrm{nd}(i)$. So either $X_i \perp X_j \mid X_{\mathrm{pa}(i)}$ or $X_j \perp X_i \mid X_{\mathrm{pa}(j)}$ by decomposition axiom. For simplicity, let $C$ be one of $X_{\mathrm{pa}(i)}$ and $X_{\mathrm{pa}(j)}$ depending on the situation. We have $X_i \perp X_j \mid X_C$. Figure 14 shows an example.

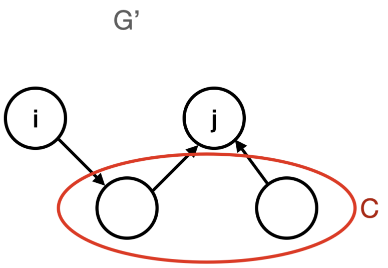


**Figure 14: An example of $\mathcal{G}'$, where $i$ and $j$ are not adjacent, but $j$ is $i$'s descendant**

If $C = \emptyset$, straight contradiction to the assumption $X_i \not\perp X_j$. Otherwise, pick $c \in C$, by assumption, $X_c \perp X_{V\setminus c}$. This is a strong argument saying $X_c$ is independent of anything else. We should be able to remove $c$ from the condition and obtain $X_i \perp X_j \,|\, X_{C\setminus\{c\}}$.

Using the graphoid axioms,

$$\begin{aligned} X_c \perp X_j \,|\, X_{C\setminus\{c\}},\ X_i \perp X_j \,|\, X_c, X_{C\setminus\{c\}} &\implies X_c, X_i \perp X_j \,|\, X_{C\setminus\{c\}} && \text{by (iv) contraction axiom} \\ &\implies X_i \perp X_j \,|\, X_{C\setminus\{c\}} && \text{by (ii) decomposition axiom} \end{aligned}$$

Left to prove $X_c \perp X_j \,|\, X_{C\setminus\{c\}}$, which is true by the following arguments:

$$\begin{aligned} X_c \perp X_{V\setminus c} &\implies X_c \perp X_j, X_{C\setminus\{c\}} && \text{by (ii) decomposition axiom} \\ &\implies X_c \perp X_j \,|\, X_{C\setminus\{c\}} && \text{by (iii) weak union axiom} \end{aligned}$$

If $C\setminus\{c\}$ is still not empty, we can keep removing nodes until arriving at the argument $X_i \perp X_j$. Contradiction. □

Can we find a classification criterion for Markov equivalence? The only exception among the local structures of three nodes in DAG seems to be the v-structure. What if the moral graphs are the same?

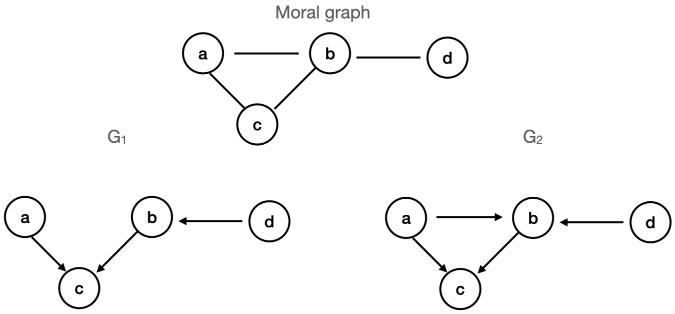


**Figure 15: Two graphs with the same moral graph but not Markov equivalent**

In figure 15, $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same moral graph. But local Markov property for $\mathcal{G}_1$ implies $X_a \perp X_b, X_d$ whereas $\mathcal{G}_2$ only has $X_a \perp X_d$ as $b$ is descendant of $a$. One can easily construct a probability $p$ s.t. $X_a \perp X_d[p]$ but $X_a \not\perp X_b$. Such $p$ is Markov on $\mathcal{G}_1$ but not on $\mathcal{G}_2$. Therefore, the moral graph cannot be used for classification.

What if we ensure all the v-structures are the same in addition to the skeletons being the same? Would Markov equivalence be implied? The answer is yes.

**Theorem 3.5.** *DAG $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent $\Leftrightarrow$ they have the same skeletons and v-structures*

*Proof.* **Markov equivalence $\Rightarrow$ same skeleton and v-structures**
Again, proving the converse (different skeleton OR different v-structure $\Rightarrow$ not Markov equivalent) is easier. The case with different skeletons is proved by the above lemma. Now, we assume $a \to c \leftarrow b$ is v-structure in $\mathcal{G}$ not in $\mathcal{G}'$. Similar to the above lemma, we cheat a bit and use $p$ satisfying

1. $X_a \perp X_b$ but $X_a \not\perp X_b \mid X_c[p]$ (this corresponds to the $v$-structure)
2. $X_v \perp X_{V \setminus \{v\}}[p]$ for all $v \notin \{a, b, c\}$

$p$ is Markov on $\mathcal{G}$, as in the moral graph $(\mathcal{G})^m$, $\{a, b, c\}$ is a complete set. So global Markov property does not require any independence between $a, b$ and $c$.

By the assumptions,

$$\begin{aligned} p(x_V) &= p(x_a, x_b)p(x_c \mid x_a, x_b) \prod_{v \in V \setminus \{a,b,c\}} p(x_v) && \text{by assumption 2} \\ &= p(x_a)p(x_b)p(x_c \mid x_a, x_b) \prod_{v \in V \setminus \{a,b,c\}} p(x_v) && \text{by assumption 1} \\ &= p(x_c \mid x_a, x_b) \prod_{v \in V \setminus \{c\}} p(x_v) \end{aligned}$$

so finding $p$ satisfying the equation ensures the two assumptions are satisfied.

Assume for contradiction that $p$ is Markov w.r.t. $\mathcal{G}'$. It is tempting to deduce from "no v-structure $a \to c \leftarrow b$ in $\mathcal{G}'$" that $a$ and $b$ are not adjoint in the moral graph $(\mathcal{G}'_A)^m$ where $A := \text{an}_{\mathcal{G}'}(\{a, b, c\})$. However, we have to ensure no $d \in A$ s.t. $a \to d \leftarrow b$ exists. Suppose, for contradiction, such $d$ exists. Note $d \in A$, so $d$ is the ancestor of $a$ or $b$ or $c$ by definition. Claim: $d$ must be an ancestor of $a$ or $b$.

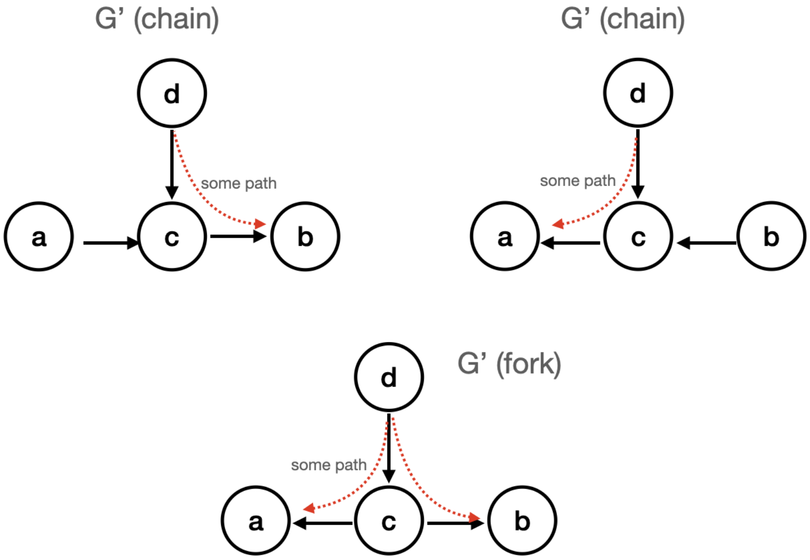


**Figure 16: Illustration of three possible local structures and the relation with $d$. The path is indicated by the red dotted arrow**

Suppose $d$ is the ancestor of $c$, listing all three possible local structures of $\{a, b, c\}$ (figure 16), we can see that the path leads to either $a$ or $b$. Claim proved.

With v-structure $a \to d \leftarrow b$, $a$ and $b$ become the ancestor of $d$. So there is a path from $a$(or $b$) to $d$ and then from $d$ to $a$(or $b$) by the claim, creating a cycle. Contradiction. So $a$ and $b$ are not involved in any v-structure.

Therefore, we can conclude that $a$ and $b$ are not adjoint in the moral graph $(\mathcal{G}'_A)^m$ and $X_a \perp X_b \,|\, X_{A\setminus\{a,b\}}$ by global Markov property. Note any element $v \in X_{A\setminus\{a,b\}}$ except $v = c$ satisfies $X_v \perp X_{V\setminus\{v\}}[p]$ by assumption, so similar removal technique in the proof of lemma can be applied, and we eventually arrives at $X_a \perp X_b \,|\, X_c$. Contradiction.

**same skeleton and v-structures $\Rightarrow$ Markov Equivalence**
The full proof can be found in [25]. □

Recall when $p > 0$ (non-degenerate distribution), two Markov properties of the graph and factorisation of $p$ are equivalent. So, Markov equivalence between undirected and directed graphs can be discussed without ambiguity. When would DAG $\mathcal{G}$ be Markov equivalent to an undirected graph $\mathcal{U}$? And what is their relation? Is $\mathcal{U}$ the moral graph or the skeleton?

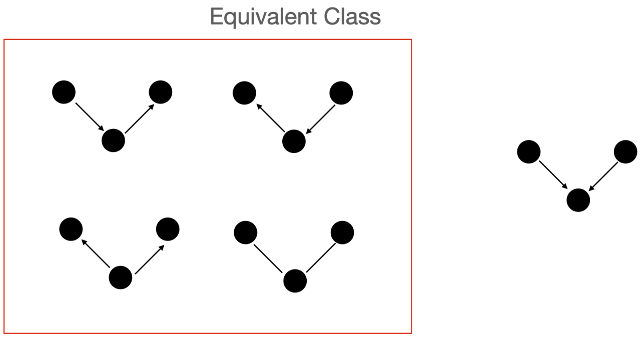


**Figure 17: Equivalent classes of the local structures**

Among the group of local structures for three nodes (with two edges), four of them (shown on the left of figure 17) are Markov equivalent to the v-structure staying alone. Also note that without v-structure, $\mathcal{G}^m = \text{skel}(\mathcal{G})$.

**Theorem 3.6.** *DAG $\mathcal{G}$ is Markov equivalent (only considering distributions with $p(x_V) > 0$) to an undirected graph iff $\mathcal{G}$ has no v-structure. And in this case, $\mathcal{G}$ is equivalent to $\mathcal{G}^m = skel(\mathcal{G})$.*

*Proof.* **Markov equivalent to UG $\Rightarrow$ no v-structure**
Suppose for contradiction that $\mathcal{G}$ has a v-structure $i \to k \leftarrow j$.
Case 1. For any undirected graph where $i \not\sim j$: pairwise Markov requires $X_i \perp X_j \,|\, X_{V\setminus\{i,j\}}\,[p]$
Note $i \sim j$ on $\mathcal{G}^m$, by the completeness of global Markov property, there is a $p$ Markov w.r.t. $\mathcal{G}$ and $X_i \not\perp X_j \,|\, X_{V\setminus\{i,j\}}\,[p]$. So, such $p$ is not Markov on any undirected graph with $i \not\sim j$.
Case 2. For any undirected graph where $i \sim j$: there is no requirement on the independence between $i$ and $j$, so distribution $p$ s.t. $X_i \not\perp X_j\,[p]$ and $X_v \perp X_{V\setminus\{v\}}\,[p]$ for any $v \neq i, j$ is Markov on this undirected graph. But local Markov property on $\mathcal{G}$ requires

$$X_i \perp X_j \,|\, X_{pa(i)}[p]$$

so $p$ is not Markov on $\mathcal{G}$.
Therefore, any undirected graph is not equivalent to $\mathcal{G}$.

**no v-structure $\Rightarrow$ Markov equivalent to UG**
Suppose $\mathcal{G}$ has no v-structure. By Proposition 3.2, factorisation passes from $\mathcal{G}$ to $\mathcal{G}^m$. So any $p$ Markov on $\mathcal{G}$ will also be Markov on $\mathcal{G}^m$. Now suppose $p$ is Markov on $\mathcal{G}^m$. Aim: show $p$ is Markov on $\mathcal{G}$.

Prove by induction on $|V|$. For simplicity, we pick a vertex $v$ with no child in $\mathcal{G}$ (why such vertex must exist?), then the additional requirement on independence after adding $v$ to $\mathcal{G}_{V\setminus\{v\}}$ is only:

$$X_v \perp X_W \,|\, X_{\text{pa}(v)}$$

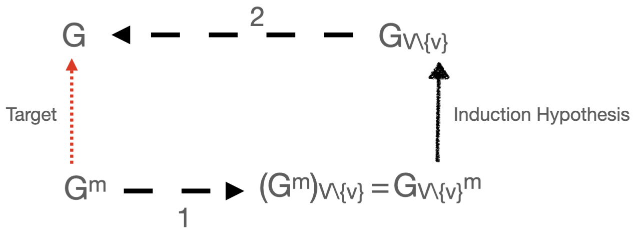


**Figure 18: Roadmap of proving if $p$ is Markov on $\mathcal{G}^m$, then $p$ is Markov on $\mathcal{G}$**

where $W := V \setminus (\{v\} \cup \mathrm{pa}(v))$. (corresponds to step 2 on figure 18) Note $\mathcal{G}_{V \setminus \{v\}}$ also has no v-structure, so $(\mathcal{G}^m)_{V \setminus \{v\}} = (\mathcal{G}_{V \setminus \{v\}})^m$. By induction hypothesis, $p$ Markov on $(\mathcal{G}_{V \setminus \{v\}})^m$ must also be Markov on $\mathcal{G}_{V \setminus \{v\}}$ So we can first prove step 1 of figure 18. $(\mathcal{G}_{V \setminus \{v\}})^m$ is the subgraph of $\mathcal{G}^m$ produced by the proper decomposition $(\{v\}, \mathrm{pa}(v), W)$. By lemma 4.23 of lecture notes, the factorisation property of $p$ on the original graph $\mathcal{G}^m$ passes to both subgraphs $(\mathcal{G}^m)_{W \cup \mathrm{pa}(v)} = (\mathcal{G}^m)_{V \setminus \{v\}}$ and $\mathcal{G}_{v \cup \mathrm{pa}(v)}$. So $p$ is also Markov on $(\mathcal{G}^m)_{V \setminus \{v\}}$. Step 1 is done.

The above decomposition also proves step 2. Because of global Markov property, decomposition $(\{v\}, \mathrm{pa}(v), W)$ implies

$$X_v \perp X_W \mid X_{\mathrm{pa}(v)} [p]$$

which brings Markov property of $p$ from $\mathcal{G}_{V \setminus \{v\}}$ to $\mathcal{G}$. □

Another natural question is: when would an undirected graph be Markov equivalent to a DAG?

**Corollary 1.** *An undirected graph $\mathcal{G}$ is Markov equivalent to a DAG iff it is decomposable*

*Proof.* **Markov equivalent to DAG ⇒ decomposable**
Suppose, for contradiction, $\mathcal{G}$ is not decomposable, then it is not triangulated and has a cycle $(c_1, c_2, \cdots, c_k)$ of length $k \geq 4$ with no chord.
If some DAG $\mathcal{G}'$ is Markov equivalent to $\mathcal{G}$, $\mathcal{G}'$ has no v-structure by the last theorem. **Claim**: $c_i \sim c_{i+1}$ for $i = 1, \cdots, k$ (let $c_{k+1} = c_1$ as this is a cycle) on $\mathcal{G}'$.
If for contradiction $c_i \not\sim c_{i+1}$ on $\mathcal{G}'$, then local Markov property on $\mathcal{G}'$ requires

$$X_{c_i} \perp X_{c_{i+1}} \mid X_A$$

for some set $A \subseteq V$ ($A$ is either $\mathrm{pa}(c_i)$ or $\mathrm{pa}(c_{i+1})$, depending on the relation between $c_i$ and $c_{i+1}$). But we can construct $p$ such that $X_{c_i} \not\perp X_{c_{i+1}} \mid X_A [p]$ while $p$ is Markov on $\mathcal{G}$ (such construction has been used many times in previous proofs). Claim proved.

Now WLOG assume $c_1 \to c_2$ on $\mathcal{G}'$, then $c_2 \to c_3$ (otherwise, we have a v-structure) This induction continues until reaching $c_k \to c_{k+1} = c_1$. But this forms a cycle on the directed graph $\mathcal{G}'$. We assumed $\mathcal{G}'$ to be acyclic. Contradiction.

**Decomposable ⇒ Markov equivalent to DAG**
Use the proper decomposition in the definition of ”decomposable graph” and prove by induction on $|V|$ (the number of vertices). □

# 4 Inference on Graphs

With all the theoretical tools about graphs and representing probabilities on graphs learnt in the last two chapters in hand, we can discuss how graphs help us with inference. Suppose the full model $p(x_V)$ is already known for all $x_V \in \mathcal{X}_V$. Statistical inference is about obtaining the probabilities $p(X_t \mid X_e)$ where $X_e$ refers to evidence and $X_t$ is the target variable to be predicted from evidence $X_e$. (This is also called *Marginal Inference*, whereas another type of inference is MAP inference, but will not be covered in this course). Three classical tools for inference are variable elimination, message passing algorithm and junction tree algorithm [29].

## 4.1 Message Passing

Firstly, we will introduce variable elimination and then generalise it to the message-passing algorithm.

The first example of chapter 7 in lecture notes is a variable elimination. But we will start with a simpler example. Consider a chain graph $x_1 \to x_2 \to \cdots \to x_n$ (so $V = \{1, 2, \cdots, n\}$), we want to find the marginal probability $p(x_n)$, the brute force algorithm is

$$\sum_{x_1} \sum_{x_2} \cdots \sum_{x_{n-1}} p(x_V)$$

But if $p$ is factorisable on the chain graph,

$$\begin{aligned} p(x_n) &= \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{n-1}} p(x_1) \prod_{i=2}^{n} p(x_i \,|\, x_{i-1}) \quad (1) \\ &= \sum_{x_{n-1}} p(x_n \,|\, x_{n-1}) \cdots \sum_{x_1} p(x_2 \,|\, x_1) p(x_1) \quad (2) \qquad \text{by rearranging the sum and product} \end{aligned}$$

This is good news for both memory and algorithm speed! Firstly, storing $p(x_V)$ for all $\mathscr{X}_V$ is not required, the probabilities $p(x_i|x_{i-1})$ (for $i = 2, \cdots, n$) plus $p(x_1)$ are enough. Note $\{i-1, i\}$ are exactly cliques on the moral graph of the chain. Secondly, (2) is less computationally complex than (1). (why? ) The example in the lecture notes also illustrates the saving of computation time by rearranging sums.

The sum (2) can be computed step by step. Firstly, replace $\phi(x_2) := \sum_{x_1} p(x_2 \,|\, x_1) p(x_1)$ (I used the symbol $\phi$ for potential function, but in this case, the sum is exactly $p(x_2)$). Then, $x_1$ is no longer in the equation. Next, $\phi(x_3) := \sum_{x_2} p(x_3 \,|\, x_2) m(x_2)$, $x_2$ is eliminated. This continues until obtaining $p(x_n) = \sum_{x_{n-1}} p(x_n \,|\, x_{n-1}) \phi(x_{n-1})$.

In general, the factorisation based on cliques can be written as

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

where $\mathcal{C}(\mathcal{G})$ is the set of cliques, $\phi$ is called potential. For the chain graph, $C = \{i-1, i\}$ and $\phi_C(x_C) = p(x_i \,|x_{i-1})$. (when $i = 1$, $\phi_i := p(x_1)$) We use potentials instead of probabilities because potential functions have no restriction of "summing up to 1". In the algorithm below, you will see operations on the potential functions like multiplication and marginalisation. Also, potential functions look cleaner than writing probabilities that involve conditioning.

**Variable Elimination Algorithm**

1. Decide an ordering of $W \subseteq V$ (variables to be eliminated) and label the variables $X_1, \cdots, X_n$ according to the order.
2. Assemble all $\phi(x_C)$ for which $X_1 \in C$, and multiply them together (call the product potential $\phi_{C^{(1)}}$, where $C^{(1)}$ is the union of all clusters containing $X_1$)
3. sum over $X_1$ (marginalisation) for $\phi_{C^{(1)}}$: i.e. define a new potential function
$$\phi_{C^{(1)} \setminus \{X_1\}} := \sum_{X_1} \phi_{C^{(1)}}$$
4. replace $\phi_{C^{(1)}}$ with $\phi_{C^{(1)} \setminus \{X_1\}}$ in the formulae. Now, $X_1$ is eliminated.
5. Repeat the above steps for $X_2, \cdots, X_n$

The choice of ordering of $W$ is essential. Some methods of choosing the ordering can be found in [3].

**Exercise 2.** *Express the "pushing summation in" example in the lecture notes (figure 19) using the potential functions described above. The ordering used is $t, b, e$. This can be seen as encoding all information of A into T, then encoding all information of S into B. Finally, push all information collected together into E. Variables X and D are children of E, while L is the child of S. So now all inferences are finished.*

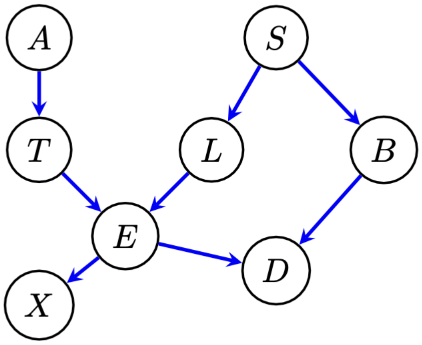


**Figure 19: The first example of Chapter 7 in lecture notes**

As explained in the exercise, variable elimination can be treated as passing a message from one node to another along the edges. In fact, if the graph can be arranged as a tree (Before turning DAG into a tree, moralise it), with target variable $x_i$ being the root and leaf nodes being evidence or the observable variables, then the post-order on a tree (an ordering where each node must be visited after all its children are visited) naturally gives an order for variable elimination. For example, in figure 20, $X_1, X_2, X_3$ are the leaf nodes(observed variables) sending messages together to the target variable $X_i$.

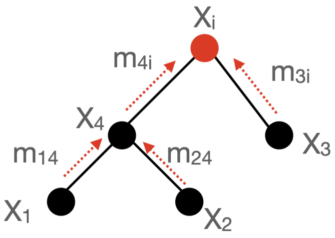


**Figure 20: An example of message-passing on a tree. The red dotted arrows are directions of message passing, but the graph is undirected**

The message passing from a leaf node $j$ to its parent $i$ is simply

$$m_{j \to i}(x_i) = m_{ji}(x_i) := \sum_{x_j} \phi(x_j)\phi(x_i, x_j)$$

where $\phi$ is the potential on a tree, defined by the following factorisation

$$p(x_V) = \prod_{i \in V} \phi(x_i) \prod_{(i,j) \in E} \phi(x_i, x_j)$$

$\phi(x_i)$ can be viewed as marginal information in $X_i$ and $\phi(x_i, x_j)$ is the interaction between $X_i$ and $X_j$. After passing the message, node $X_j$ is ”removed” in variable elimination.

The general definition of the message being passed from any node $j$ to node $i$ is

$$m_{ji}(x_i) = \sum_{x_j} \phi(x_j)\phi(x_i, x_j) \prod_{n \in N(j) \setminus \{j\}} m_{nj}(x_i)$$

where $N(j)$ is the set of neighbours of $X_j$. You can also view this as

$$m_{ji}(x_i) = \sum_{x_j} \phi'(x_j)\phi(x_i, x_j) \tag{4}$$

where $\phi'(x_j)$ is the updated version of $\phi(x_j)$ up to this step defined by

$$\phi'(x_j) := \phi(x_j) \prod_{n \in N(j)\setminus\{j\}} m_{nj}(x_i) \tag{5}$$

i.e. $\phi'$ has already stored all the information from $X_j$'s children.

The advantage of message passing is that, suppose we changed our minds and decide to compute the marginal probability of another variable $X_{\text{alt}}$, then compute

$$\beta(x_{\text{alt}}) := \phi(x_{\text{alt}}) \sum_{n \in N(j)} m_{n \to \text{alt}}(x_{\text{alt}})$$

i.e. update potential by multiplying the sum of messages from all neighbours. $\beta(x_{\text{alt}})$ refers to belief (this algorithm is also called *belief propagation*) and it is proportional to $p(x_{\text{alt}})$. So $p(x_{\text{alt}})$ can be found by normalisation

$$p(x_{\text{alt}}) = \frac{\beta(x_{\text{alt}})}{\sum_{x_{\text{alt}} \in \mathscr{X}_{\text{alt}}} \beta(x_{\text{alt}})}$$

## 4.2 Junction Trees

Not every graph can be arranged as a tree. For example, figure 21 is not a tree because $X_1$ is stuck in a 3-cycle.

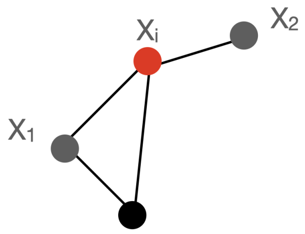


**Figure 21: A graph that is not a tree. Grey nodes are the observed variables and black nodes are unobserved**

Another problem is that the variables usually come in groups. For example, if $C_e$ is the group of evidence and $C_t$ is a group of variables to be predicted. We do not care about relations between evidences, but how all evidence affects the target variables together. Therefore, we can shrink each group/cluster into a single node. And an edge represents the intersection between two groups. This was the idea of clustering graph introduced in Chapter 2.

Now define the *junction graph* similarly to the clustering graph. But there is no need to add an edge between every two clusters with an intersection. The choice of edges can be arbitrary.

A junction graph that can legally perform message passing (or variable elimination) is called a *junction tree*. Rigorously, we require, for all paths between $C_i$ and $C_j$ on the cluster graph, each cluster on the path contains $C_i \cap C_j$. Otherwise, imagine in figure 22, a message from $X_1$ is trying to pass to $C_3$ through $C_2$, but $X_1$ is not even in $C_2$. The message-passing formula fails.

Recall the intuition of RIP from the view of the cluster graph. Elements in $C_i$ that appeared in the history are contained in a single cluster $C_{\sigma(i)}$ in the history. i.e. $C_i$ is like a leaf attached to the existing tree built from the

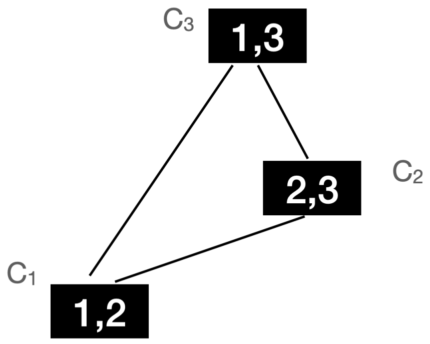
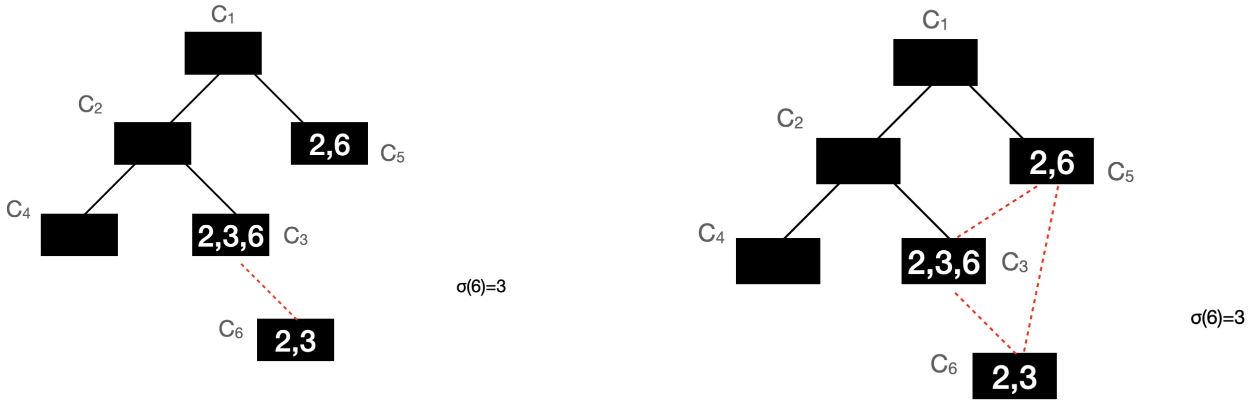


**Figure 22: A junction graph that is not junction tree**

first $i-1$ clusters. (see an example in figure 23a) It is possible for $C_i$ to have intersection with other historical set. For example, in figure 23b, $C_6$ has an intersection with both $C_3$ and $C_5$. However, the intersection with $C_5$ will be contained in $C_3$. The way $C_6$ is attached to the tree ensures its path to any historical set is either direct or through $C_5$ first. So the edge between $C_5$ and $C_6$ is unnecessary. We will remove it to build a simpler junction tree. (like figure 23a)



**(a) attaching a leaf $C_6$ to the existing tree, according to the fact that $C_6 \cap H_6 = C_6 \cap C_3$. (i.e. $\sigma(6) = 3$)**

**(b) example of a cluster with multiple connections to the history**

**Figure 23: Building Junction tree by attaching the next cluster in RIP order as a leaf**

From the view of variable elimination, removing the new leaf $C_6$ will not affect any paths between previous clusters. Because $C_6$ is somehow shielded by $C_3$, all historical information is in $C_3$, so if any information(node) from cluster $C_6$ is required, going through $C_3$ is enough. There is no need to take a long way around through $C_6$.

**Proposition 4.1.** *A cluster graph built from $C_1, \cdots, C_k$ is a junction tree iff the clusters have an ordering satisfying RIP.*

*Proof.* See rigorous proof in lecture notes proposition 7.2 □

For the message-passing algorithm, we need to define potentials. For cluster $C$, assign potential $\phi_C(x_C) \geq 0$. Also, assign potentials $\phi_S(x_S)$ to each edge of the junction tree. (i.e. potentials for the separator sets $S$) These will be used to hold messages. (similar to figure 20, where messages are assigned to edges) The key difference between the previous algorithm (where the graph can be arranged into a tree) is that potentials in the previous algorithm are assigned to variable $X_i$ directly. So, the probability distribution $p(x_i)$ can be obtained by normalisation. Converting $\phi_C(x_C)$ to $p(x_C)$ can also be done by normalisation, but the concept of consistency makes things easier.

If clusters $C$ and $D$ are adjacent on the junction tree $\mathcal{T}$, $\phi_C$ and $\phi_D$ are called *consistent* if they carry the same piece of information on the intersection $S := C \cap D$.

**Definition 14.** $\phi_C$ and $\phi_D$ are consistent for clusters $C, D$ if

$$\sum_{x_{C \setminus D}} \phi_C(x_C) = f(x_S) = \sum_{x_{D \setminus C}} \phi_D(x_D)$$

i.e. margins of two potential functions on $S$ are the same.

If $\phi_C(x_C) = p(x_C), \phi_D(x_D) = p(x_D)$ are probability distributions, then by the law of total probability, they must be consistent. Does consistency imply potential functions are probability functions? It holds on decomposable graphs.

**Proposition 4.2.** *Let $C_1, \cdots, C_k$ be an ordering of cliques satisfying RIP with separator sets $S_2, \cdots, S_k$. Suppose*

$$p(x_V) = \prod_{i=1}^{k} \frac{\phi_{C_i}(x_{C_i})}{\phi_{S_i}(x_{S_i})} \tag{6}$$

*then*

$$\phi_{C_i}(x_{C_i}) = p(x_{C_i}),\ \phi_{S_i}(x_{S_i}) = p(x_{S_i}) \ \Leftrightarrow\ \textit{all pairs of } \phi_{C_i} \textit{ and } \phi_{S_i} \textit{ are consistent}$$

*Proof.* $\Rightarrow$ direction is already proved in the comments above.
For $\Leftarrow$, our old friend induction helps. RIP means $S_k := C_k \cap H_k = C_k \cap C_{\sigma(k)}$ for some $\sigma(k) < k$, then make use of consistency of potentials to prove $p(x_{S_k}) = \phi_{S_k}(x_{S_k})$. The other equation $\phi_{C_k}(x_{C_k}) = p(x_{C_k})$ can be proved using the factorisation formulae. □

So if message passing ensures all potentials are consistent and the potential factorisation equation 6 holds, then the potentials obtained are exactly the marginal probabilities we queried for inference.

Message passing from $C$ to $D$ for junction trees is done by

$$m_{D \to C}(x_S) := \sum_{x_{C \setminus S}} \phi_C(x_C)$$

where $S := C \cap D$. For convenience denote $m_{D \to C}(x_S)$ as $\phi'_S(x_S)$ instead. i.e. update the potential on the separator (edge of junction tree) to store the message. $D$ receives the message, so update $\phi_D$ by

$$\phi'_D(x_D) = \frac{\phi'_S(x_S)}{\phi_S(x_S)} \phi_D(x_D)$$

Note how these two equations are similar to equations 4 (message) and 5 (potential update). Correspondence:

- Interaction: $1/\phi_S(x_S)$ and $\phi(x_i, x_j)$
- Message: $\phi'_S(x_S)$ and $m_{ji}(x_i)$
- Margin: $\phi_D(x_D)$ and $\phi(x_i)$, $\phi_C(x_C)$ and $\phi(x_j)$

How to use message passing to make all potentials consistent?

Note $\phi_C$ is consistent with $\phi'_S$ (defined as margin of $\phi_C$) by definition. If $\phi_D$ and $\phi_S$ were consistent, this message passing would not break the consistency as we scaled $\phi_D$ according to the margin by $\phi'_S/\phi_S$. i.e. $\phi'_D$ and $\phi'_S$ are also consistent. Suppose we then pass a message from $D$ back to $C$, obtaining $\phi''_C(x_C)$ and $\phi''_S(x_S)$. Then

$$\phi''_C(x_C)$$

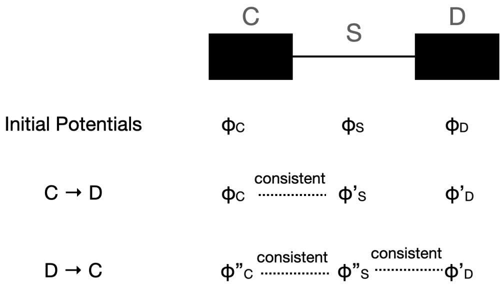


**Figure 24: Message passing on a junction tree with only two clusters**

is consistent with $\phi''_S$ as $\phi_C$ and $\phi'_S$ are consistent. And second passing makes $\phi''_S$ consistent with $\phi'_D$. (See figure 24 for illustration) Now the final round potentials are all consistent.

Note also the product of potentials in equation 6 remains constant. By definition of the new potential

$$\frac{\phi'_D(x_D)}{\phi'_S(x_S)} = \frac{\phi_D(x_D)}{\phi_S(x_S)}$$

So as long as the initial potentials satisfy equation 6, all updated potentials satisfy this equation.

**Junction Tree Message Passing Algorithm**: For a general junction tree, pass messages from leaf nodes to the root first, then pass message pack to leaf nodes. The two steps are called collection and distribution. By similar arguments to the above, all pairs of potentials are consistent after the algorithm.

**Theorem 4.3.** *For a decomposable graph $\mathcal{G}$ with clique ordering $C_1, \cdots, C_k$ satisfying RIP, let the initial choice of potential on any clique $C$ be*

$$\phi_C(x_C) := \prod_{v \in v(C)} p(x_v \mid x_{pa(v)})$$

*where $v(C)$ are vertices in cluster $C$. And let $\phi_S(x_S) = 1$.*

*After the collection and distribution steps of the junction tree message passing algorithm, all potentials become consistent and equation 6 still holds. Hence, updated potentials are exactly the marginal probabilities of the clusters.*

The theorem says with appropriate initial choice, the junction tree message passing algorithm produces correct marginal probabilities.

*Remark.* For non-decomposable graphs, finding RIP ordering is not possible. Loopy Belief Propagation (LBP) is one of the approximate inference techniques designed for such models. It is a special case of variational inference. Another type of approximate inference technique is sampling-based inference(like MCMC simulation) A good introduction can be found on https://ermongroup.github.io/cs228-notes/inference/sampling/.
Non-decomposable graphs can also be turned decomposable by adding edges to make the graph triangulated. There are many ways to triangulate, but the resulting graph with smaller cliques $C_i$ is preferable, which reduces computation times of marginalisation in message passing.

*Remark.* Whenever there is a change in the model, for example, having the evidence that a variable $X_i = x$ for a known $x$. Then we can modify the potential function accordingly (concentrating potentials of all $C$ containing $X_i$ at value $x$ ), and then propagate the message to all other nodes. So inference using message passing is very flexible.

# 5 Causal Inference

Knowledge is the object of our inquiry, and men do not think they know a thing till they have grasped the 'why' of it (which is to grasp its primary cause) —— *Physics*, Aristotle

Philosophers have been thinking about what "causation" is for thousands of years. Our beliefs about causation depend on experience, but are they reliable? For example, seeing many people curing their colds fast after taking vitamin C, you believe that vitamin C causes colds to cure. What if it is the mental comfort after taking vitamin C cures a cold? In David Hume's idea, experience can only be used to verify *constant conjunction* (similar to covariance in statistics). The detailed philosophical background of "causation" can be found in chapter 2 of Functorial Causal Models: Towards an Algebraic Approach to Causal Inference. Indeed, verifying causal relations can be hard. Works by great statisticians like Neyman, Rubin and Pearl give us tools to express causation in mathematical language.

Pearl introduced graphical tools to help with causal relations [13], and the lecture notes mainly focus on his model. This note is largely based on Pearl's book[14] and another book by Hernan and Robins[7].

Pearl's Causation Hierarchy [15] summarises three levels of our knowledge: association, intervention, and counterfactual.

- **Association**: concluded from observations. If $X$ and $Y$ occur together frequently, seeing $X$ makes you believe $Y$ is going to happen. e.g. $X$ - symptoms of cold, $Y$ - you have a cold.
- **Intervention**: Doing interventions. What if I do X? e.g. would taking vitamin C cure my cold?
- **Counterfactual**: Imagining and understanding. Was it $X$ causing $Y$? What if I take a different action? e.g. what if I did not take vitamin C? Would taking vitamin D also cure my cold?

Neyman [22] proposed a mathematical model to measure the causal effect of treatment on the outcome, and then Rubins refined it for observational experiments. For simplicity, assume there are only two treatments $a = 0$ and $a = 1$ (not treat and treat). $Y^{a=0}$, $Y^{a=1}$ are called *potential outcomes*, where $Y^{a=0}$ means the potential outcome when not treated and $Y^{a=1}$ is the potential outcome when treated. They are called 'potential' because, in a real experiment, you can only observe one of them (see the example below). The actual treatment in the experiment is denoted $A$, and the actual outcome is $Y$.

**Example 5.** You wonder whether a pill kills a mouse. $a = 0$: not take the pill, $a = 1$: take the pill, $Y^{a=0}$: death if not given the pill, $Y^{a=1}$: death if given the pill. Suppose you gave the mouse a pill ($A = 1$) and it dies($Y = 1$). You cannot revive it and experiment again without giving the pill.

**Definition 15** (Consistency)**.** The actual outcome is the same as the potential outcome corresponding to the treatment given. i.e. when $A = a$, we have $Y = Y^a$.

*Remark.* Consistency seems obvious, but the actual treatment ($A$) may take a different form than what you proposed ($a$). For example, doctors may do the same surgery in slightly different ways. Consistency is saying your proposed treatment and actual treatment always result in the same outcome.

**Definition 16** (Individual Causal Effect)**.** Treatment $a$ has causal effect on $Y$ if

$$Y^{a=1} \neq Y^{a=0}$$

i.e. treating ($a = 1$) makes a difference. The magnitude of the causal effect is $Y^{a=1} - Y^{a=0}$. If the treatment is not binary, $a$ has a causal effect on $Y$ if there are treatments $a_1, a_2$ s.t.

$$Y^{a=a_1} \neq Y^{a=a_2}$$

For an individual, checking the causal effect is impossible. Because among all potential outcomes, you can only observe the one corresponding to the actual treatment. In an experiment, a population is taken, and $A_i, Y_i$ stands for the actual treatment and outcome of individual $i$.

**Definition 17** (Average(population) Causal Effect)**.** Treatment $a$ has a causal effect on binary outcome $Y$ in a population if

$$P(Y^{a=1} = 1) \neq P(Y^{a=0} = 1)$$

In general, $a$ has a causal effect on outcome $Y$ if

$$E(Y^{a=1}) \neq E(Y^{a=0})$$

This definition also works for continuous $Y$.
The generalisation of non-binary treatment is similar to the previous definition.

*Remark.* Note if $E(Y^{a=1}) \neq E(Y^{a=0})$, by linearity,

$$E(Y^{a=1} - Y^{a=0}) \neq 0$$

where $E(Y^{a=1} - Y^{a=0})$ represents the magnitude of causal effect. The inequality means the average individual causal effect $Y_i^{a=1} - Y_i^{a=1}$ across the population is non-zero.

The data you collect for the population could take the form below (binary treatment and outcome)

| individual $i$ | treatment $A_i$ | outcome $Y_i$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 0 |
| 3 | 0 | 0 |
| 4 | 1 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 0 |

Assuming consistency, we can add columns representing the potential outcomes.

| individual $i$ | treatment $A_i$ | outcome $Y_i$ | $Y_i^{a=0}$ | $Y_i^{a=1}$ |
|---|---|---|---|---|
| 1 | 1 | 1 | ? | 1 |
| 2 | 1 | 0 | ? | 0 |
| 3 | 0 | 0 | 0 | ? |
| 4 | 1 | 1 | ? | 1 |
| 5 | 0 | 1 | 1 | ? |
| 6 | 0 | 0 | 0 | ? |

where the question marks are there because the outcome for that individual is not observed. If the actual treatment is 0, we say $Y_i^{a=1}$ is the outcome in the counterfactual world where individual $i$ was treated, and vice versa. So potential outcomes are also called *counterfactual outcomes*.

If the experiment is conducted randomly, i.e. you randomly assign the treatment $A_i$ to individuals by flipping a coin (does not need to be fair), then the causal effect equals the correlation difference, i.e.

$$P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = P(Y = 1 \mid A = 1) - P(Y = 1 \mid A = 0)$$

the reason will be explained later. But if you have observational data, the situation is different.
**Causation $\neq$ Correlation**. The main difference is that causation is unconditional, it compares two counterfactual worlds where in one the whole population is treated, and not treated in the other. Correlation compares two subsets (treated and untreated) of a population in the factual world. See a graphical illustration in figure 25 where the square represents the population.

In randomised experiments, causation = correlation. Because *exchangeability* holds. Suppose you wish the individual to be treated if the coin flipped to the head and untreated if the tail, but your assistant mistakenly treated the group with the coin flipped to the tail instead. The data is still valid, as we assume the potential outcome $Y^a$ to be an inherent quantity, not affected by the actual treatment.

**Definition 18** ((full) Exchangeability)**.** If $Y^a \perp A$, we say exchangeability holds.

If exchangeability holds,

$$P(Y^a = 1) = P(Y^a = 1 \mid A = 1) = P(Y^a = 1 \mid A = 0)$$

so the causal effect equals the difference in conditional probabilities (association/correlation). The right side of the equation can be calculated directly from the data. In fact, the mean exchangeability ($E(Y|A = a) = E(Y^a|A = a) = E(Y^a)$) is enough for association = causation.

Consider the following setting: $L$ - body condition, $A$ - treatment, $Y$ - death. You doubt that the causal effect may be different for patients in good condition ($L = 0$) and critical condition ($L = 1$). In such cases, you may flip two coins to decide the treatment $A$. One for $L = 1$, one for $L = 0$. For example, give 50% of population $L = 0$ and 75% of $L = 1$ treatment. The experiment is said to be *conditionally randomised*. Exchangeability does not hold in this case, because $A$ depends on $L$. But conditional exchangeability holds.

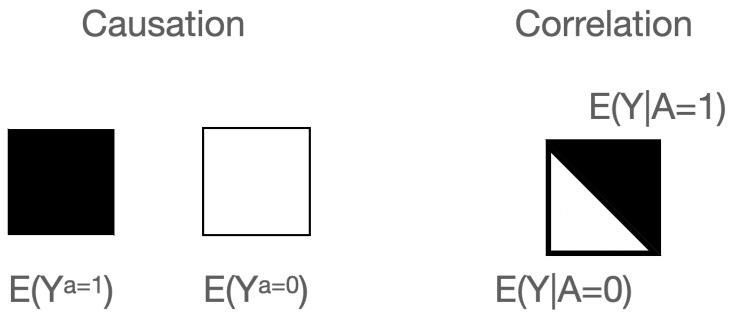


**Figure 25: Causation vs Correlation**

**Definition 19** (conditional exchangeability)**.** $Y^a \perp A \,|\, L$. i.e. within each level of $L$, counter-factual $Y^a$ is independent of the actual treatment $A$

With conditional exchangeability, the conditional causal effect equals the conditional correlation difference, i.e.

$$P(Y^{a=1} = 1 \,|\, L = l) - P(Y^{a=0} = 1 \,|\, L = l) = P(Y = 1 \,|\, a = 1, L = l) - P(Y = 1 \,|\, a = 0, L = l)$$

for all levels $l$.

**Definition 20** (effect modification)**.** If the conditional causal effect

$$P(Y^{a=1} = 1 \,|\, L) - P(Y^{a=0} = 1 \,|\, L)$$

for different values of $L$ are different, we say $L$ modifies the causal effect of $A$ on $Y$. WARNING: this does not mean $L$ causes $A$ or $L$ causes $Y$.

To calculate the population causal effect from conditional causal effects, use the law of total probability,

$$P(Y^a = 1) = \sum_l P(Y^a = 1 \,|\, L = l) P(L = l)$$

where $P(L = l)$ is the portion of $l$ in the population. This method is called *standardisation.*

## 5.1 Observational Study and Rubin's model

In the introduction part, we explained how randomised and conditionally randomised experiments allow you to calculate causal effects from the data. However, most experiments do not allow ideal randomisation (e.g. study whether smoking causes cancer. Experimenters cannot force people to smoke nor stop people from smoking). Further, many data collected are from observations (evolutionary studies, fossils).

Rubin proposed some criteria called *identifiability* (i.e. you can identify causal effect from the data).

- Consistency: treatments $a$ well-defined, and $Y^a = Y$ when $A = a$.
- Exchangeability/no-confounding (Rubins called it *ignorability*): conditional probabilities only depend on the measured covariates/factors $L$. (In observational studies, any variables other than treatment $A$ and outcome $Y$ are called *covariates/factors*)
- Positivity: each treatment appears at least once in each level of $L$. i.e. $A|L$ has a positive distribution

The concept of *effect modification* can also be used for observational studies. A covariate $V$ modifies the effect if the causal effect of $A$ on $Y$ varies across levels of $V$. Studying effect modification is important because

- Trans-portability: with effect modification by $V$, the measured causal effect cannot be transferred to another population unless the distribution of $V$ is the same in two populations.
- We can identify which group benefits the most from the treatment. e.g. if $V$ is sex, taking $V$ into account identifies whether males and/or females benefit from the treatment. (Also see the example below)

**Example 6.** Suppose a treatment increases the risk of death for males, but decreases the risk for females. Since the causal effects go opposite directions, the population causal effect may be cancelled. If there are more females in the sample, the conclusion could be treatment reduces the risk of death. So if the experimenter did not take covariate $V$(sex) into account, it could harm the individuals at some levels of $V$. This is one case of *Simpson's paradox*. (see figure 26)

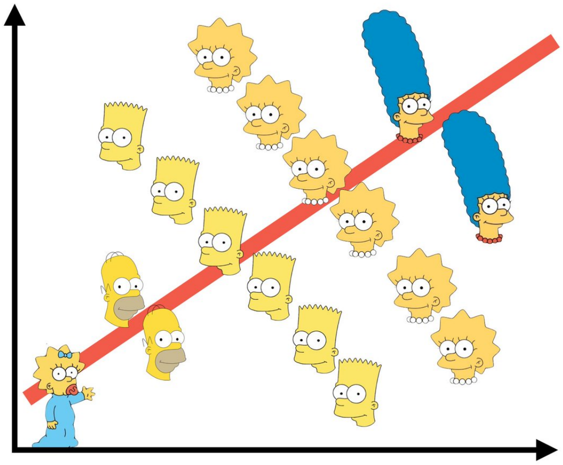


**Figure 26: Simpson paradox with Simpson family. (presented by Instagram @infowetrust)**

*Adjustments* are required for effect modifications. One strategy is *Stratification*, as many covariates (say $L = (L_1, \cdots, L_k)$ are covariates) are measured as possible, and causal effects are considered on each stratum (level) of $L$. Then, standardisation (summing all conditional probabilities using the law of total probability) can be used to find the population causal effect. This adjustment method is used in the lecture notes. Also, Rubin proposed propensity score (which allows inverse probability weighting(IPW) for adjustment) in [20] and matching in [19]. (this method will not be introduced here)

It is worth mentioning that Rubin's model does not require graphs. However, the ignorability/exchangeability assumption may not hold in many experiments, as it is hard to ensure all causes of outcome are identified and measured. Also, the model assumes $Y$ appears later than $A$, but they may be produced simultaneously.

## 5.2 Causal Diagrams and Pearl's model

To aid data analysis and present causal relations better, Pearl used graphs (DAG) [13]. Instead of talking about counterfactuals, Pearl used the 'do' operator on the graph for interventions. Figure 27 shows two examples of causal diagrams. Variables in causal diagrams are usually ordered in time. Variables to the left happen earlier.

**Definition 21** (Causal DAG)**.** A DAG with vertices $V = \{V_1, \cdots, V_M\}$ is a causal diagram if

- (1) absence of edge from $V_i \to V_j$ implies no direct causal effect of $V_i$ to $V_j$
- (2) All common causes(measured and unmeasured) of any pair of vertices should be in the graph
- (3) Any variable is a cause of its descendants. (But direct causes of $v$ are its parents $\text{pa}(v)$)

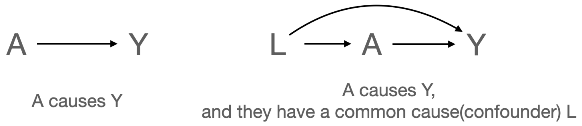


**Figure 27: Examples of causal diagrams**

- (4) Causal Markov assumption: Node $V_j$ is independent of any $V_i$(not caused by $V_j$) conditional on direct causes of $V_j$. (note, this is exactly the local Markov property for DAG)

Assumption (4) implies (2) because if there is any hidden common cause $L$, it makes $A$ and $Y$ correlated ($A \not\perp Y$). But on the graph without $L$, the Markov assumption requires $A \perp Y$.

There is a counterfactual model underlying causal DAG called FRCISTG(Fully Randomised Causally Interpreted Structured Tree Graph) [18]. It builds counterfactuals recursively by assigning (unknown) functions

$$f_m(\text{pa}(V_m), \epsilon_m)$$

to counterfactual $V_m^{v_1,\cdots,v_{m-1}}$, i.e. assumes all causal effects on $V_m$ come from its parents. $\epsilon_m$ is a random variable representing error. Such models are called *structural equation models* (SEM).(more details of SEM in section 5.7) DAG representing a FRCISTG has a factorisation of joint PDF $p(x_V)$:

$$p(x_V) = \prod_{i=1}^{M} f(x_i \,|\, \text{pa}(V_i)) \quad (*)$$

for some function $f$, where $x_i := x_{V_i}$. Any causal DAG $\mathcal{G}$ with joint density $f$ satisfying equation (*) is called *causal Bayesian network*, denoted $(\mathcal{G}, f)$.

Pearl's model is very similar to FRCISTG but assumes independence between counterfactuals. Pearl's model is called NPSEM-IE(Non-Parametric Structural Equation Model with Independent Errors) [14]. Also, counterfactuals can be facilitated into causal DAG, the resulting graph is called SWIG(single-world intervention graph)[17].

The power of graphs is that: they simultaneously represent causation and association(statistical dependence) Pearl's model still has some drawbacks:

- Requires prior knowledge of causal relations to draw the causal diagram. There is a field of research called *causal identification.*
- Magnitude and direction of causal effects are not represented
- Some graphs are indistinguishable from the data. Because the graphs result in the same probability distribution. (incompleteness of Markov property)

Before studying how Pearl defines causation by intervention, let's see how association flows along the graph through some examples. Note association is a symmetric relationship, so it is not restricted by the direction of arrows. (directed edges) Causal paths, on the other hand, must flow along the arrows.

### Marginal Association

**Example 7.** Consider a simple graph with only $A \to Y$. Since there is no other vertex, $A, Y$ has no common causes. So all associations between $A$ and $Y$ are brought by the causation $A \to Y$. i.e.

$$P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = P(Y = 1|A = 1) - P(Y = 1|A = 0)$$

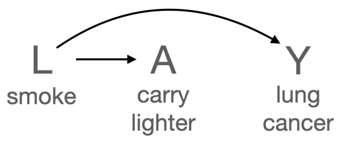


**Figure 28: Example of marginal association(confounding)**

**Example 8.** In figure 28, whether someone is carrying a lighter ($A = 1$) or not ($A = 0$) does not cause lung cancer $Y$. So there is no arrow from $A$ to $Y$. However, research data may show an association: people carrying lighters are more likely to get lung cancer. Because there is a common cause of lung cancer and carrying a lighter: smoking $L$. This is an example of no causation (i.e. $P(Y^{a=1} = 1) = P(Y^{a=0} = 1)$) but association, i.e. $P(Y = 1|A = 1) \neq P(Y = 1|A = 0)$. We say association flows through this *backdoor path* $A \leftarrow L \rightarrow Y$. (The direction of edges/arrows on the backdoor path can have opposite directions) Such a case is called *confounding* and will be studied more later.

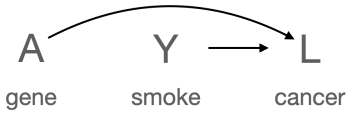


**Figure 29: Example of no marginal association(common effect)**

**Example 9.** In figure 29, $A$ is a gene haplotype that can prevent cancer. We believe that gene $A$ does not cause someone to smoke (no arrow from $A$ to $Y$). Is there an association? $A$ and $Y$ are not correlated in this case, as they are separate mechanisms affecting the risk of lung cancer. Even if the association tries to flow through the backdoor path $A \rightarrow L \leftarrow Y$, a future event should not affect how the past events $A, Y$ are related. We say the path is *blocked* by $L$. This is an example of no causation (i.e. $P(Y^{a=1} = 1) = P(Y^{a=0} = 1)$) and no association, i.e. $P(Y = 1|A = 1) = P(Y = 1|A = 0)$. This example also shows the **non-transitivity** of association. Because associations flow along $A \rightarrow L$, $Y \rightarrow L$, meaning $\{A, L\}$ and $\{Y, L\}$ are correlated. But $\{A, Y\}$ are not correlated.

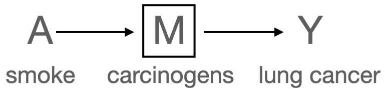


**Figure 30: Example of conditional association (chain)**

**Example 10.** Smoking causes lung cancer because it fills the lungs with carcinogens (cancer-cause substances). $M = 1$: carcinogens found in lungs, $M = 0$: no carcinogens in lungs. The causal relations are shown in figure 30. Suppose we condition on $M = 1$, i.e. already found carcinogens in lungs. Would smoking $A$ be correlated to lung cancer $Y$ through this route $A \rightarrow B \rightarrow Y$? No, because once carcinogens are found, smoking or not does not matter. Carcinogens directly cause cancer. In situations like this, $M$ is called a mediator. Conditioning on $M$ blocks (square box around $M$ means conditioned) the association flowing from $A$ to $Y$ by blocking the path $A \rightarrow B \rightarrow Y$.

**Example 11.** Examine the lighter-carrying example again(figure 31). Suppose now we only study the non-smokers ($L = 0$), then carrying a lighter is not relevant to lung cancer. If we only look at smokers $L = 1$, $A, Y$ are still

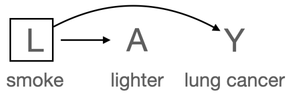


**Figure 31: Example of conditional association(confounding)**

irrelevant as lung cancer is caused by smoking. Among smokers, carrying a lighter may be a personal habit, some smokers may borrow other's lighter or smoke after returning home, so they do not carry a lighter ($A = 0$). So conditioning on the common cause $L$ blocks association by blocking path $A \leftarrow L \rightarrow Y$.

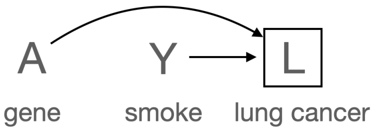


**Figure 32: Example of conditional association(common effect)**

**Example 12.** Review the gene haplotype example again (figure 32). Conditioning on $L = 1$ (i.e. only study people with cancer). Suppose $A = 0$ (no gene reducing cancer risk), they become vulnerable to other cancer-causing cases as well, so the likelihood of smoking $Y$ is lower than the group $A = 1$. i.e. $A$ and $Y$ are associated.

$$P(Y = 1 \mid A = 0, L = 1) \neq P(Y = 1 \mid A = 1, L = 1)$$

In this case, we say conditioning on the common effect $L$ opens the path $A \rightarrow L \leftarrow Y$ that was blocked by $L$ marginally. So $A$ and $Y$ become associated through the path.

Interestingly, conditioning on any descendant of $L$ also gives association. For example, in figure 33, $C$ is a consequence of lung cancer. Conditioning on $C$ provides information on $L$ and opens the path $A \rightarrow L \leftarrow Y$. The two examples introduced are *selection bias* and will be discussed later.

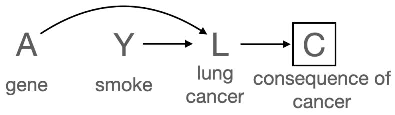


**Figure 33: Example of conditional association(descendant of common effect)**

Among all structures, only the v-structure $a \rightarrow t \leftarrow b$ behaves differently. For a causal graph, this means $t$ is the common effect of $a, b$. We define such $t$ to be a *collider*. Other edge structures are called *non-colliders*

In conclusion, on a causal DAG, there is an association between $A$ and $Y$ if there is an open path. Where *open* means

- all non-colliders on the path are not being conditioned
- all colliders or at least one of their descendants are conditioned

A path that is not open is called *blocked*.

**Definition 22** (open path)**.** For path $\pi$ from $a$ to $b$, $\pi$ is open conditional on $C \subseteq V \setminus \{a, b\}$ if

- all colliders are in $C$ or descendant of a vertex in $C$. i.e. colliders in $\text{an}_{\mathcal{G}}(C)$
- all non-colliders not in $C$

**Definition 23** (d-separation)**.** Points $a$, and $b$ are d-separated given $C$ if all paths are blocked conditional on $C$. Sets $A, B$ are d-separated if each pair of points $a \in A, b \in B$ are d-separated.

The following theorem [12] formalise the concept of: association flows along open paths that we have been talking about.

**Theorem 5.1.** *For causal DAG $\mathcal{G}$, if disjoint sets $A, B, C$ satisfies $A$ d-separated from $B$ by $C$, then $A \perp B \,|\, C$.*

The converse of this theorem (called *faithfulness*), independence ⇒ d-separation, may not hold in some cases. Suppose $A \to Y$, and $V$ modifies the effect of $A$ on $Y$ and different levels of $V$ exactly cancel out, then $A \perp Y$. But clearly, $A$ and $Y$ are not d-separated. Such occasions are rare in observational studies and most experiments (unless you deliberately make the distribution of $V$ the same across untreated and treated populations, e.g. by matching), so usually faithfulness is assumed.

Lauritzen proposed an alternative rule for separation using moral graphs [10].

**Theorem 5.2** (Equivalent separation rule on moral graph)**.** *For DAG $\mathcal{G}$ and disjoint subsets $A, B, C$. $A$ is d-separated from $B$ by $C \Leftrightarrow A$ is separated from $B$ by $C$ in $(\mathcal{G}_{an(A \cup B \cup C)})^m$.*

The above structural study shows that collecting as many covariates as possible, and conditioning on all of them may not always be a good idea. Because if the covariate happens to be a collider, additional association or *bias* is introduced (even if there is no causal relation). Bias disguises you when concluding causal relations using data. More on bias later.

## 5.3 Intervention and Adjustments

An intervention is like setting treatment $a = 1$ (or $a = 0$) and jumping into the counterfactual world. The way to represent it on the graph is by removing all parental edges of $X_a$. Because the value of $a$ is fixed, all its causes no longer have causal effects. The new graph is called *mutilated.* An intervention is denoted as $do(x_a)$, i.e. setting treatment $X_a$ to $x_a$. Denote probability distribution assigned to mutilated graph $P_m(Y|X_a = x_a) := P(Y \mid do(x_a))$, it represents the probability distribution of $Y$ after all individual's treatment is changed to $x_a$ (this is done in imagination). The causal effect of $A$ on $Y$ is defined as a function from $\mathscr{X}_a$ (possible values of $A$) to the space of all probability distributions on $Y$, $x_a \mapsto p(y|do(X_a = x_a))$. As long as this function is constant, there is a population causal effect.

But the conditional probability $P(Y|A = a)$ is not imagining a new world, it takes a subset of the factual world and studies the distribution of $Y$ on the subset. So in general, $P(Y \mid do(x_a)) \neq P(Y|A = a)$.

*Remark.* Pearl's model and Rubin's model both reveal that causal relations are too complicated to be represented only using simple probabilities. Graphs or counterfactuals $Y^a$ are required.

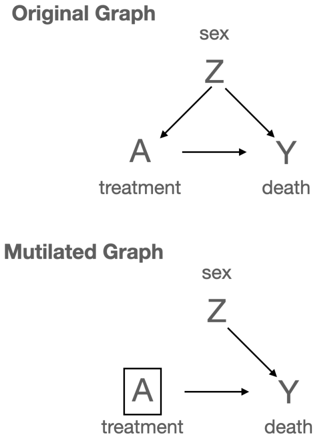


**Figure 34: Example of mutilated graph (confounding)**

**Key question**: How to calculate the intervened probability distribution $P_m(Y|X_a = x_a)$?

Take the simple example in figure 34 where the effect of treatment $A$ on death $Y$ is studied. Sex $Z$ is a common cause of both treatment and death. Two assumptions should be made

- Variable $Z$ is not a descendant of $A$, so it is not affected by intervention. i.e. $P_m(z) = P(z)$. (in this example, it encodes treatment will not affect sex)
- the response $y$ at each $z$ remains the same, i.e. $P_m(y|a, z) = P(y|a, z)$. With these two assumptions (the correlation between $A, Y$ for males(or females) remains the same in the intervened world)

$$\begin{aligned}
P(y|do(a)) &= P_m(y|a) \\
&= \sum_z P_m(y, z|a) \quad \text{Law of total probability} \\
&= \sum_z P_m(y|a, z)P_m(z|a) \\
&= \sum_z P_m(y|a, z)P_m(z) \quad \text{because } A, Z \text{ are d-separated in mutilated graph} \\
&= \sum_z P(y|a, z)P(z) \quad \text{by the above assumptions}
\end{aligned}$$

And in general, $P_m(y|a) \neq P(y|a) = \sum_z P(y|a, z)P(z|a)$. Because $Z$ and $A$ are not separated in the original graph.

So the intervened counterfactual world can be evaluated using data collected in the factual world by summing over $z$, which is a form of adjustment (adjusting the effect of $Z$ on $Y$). It can be proved that summing over all parents of $A$ (standardisation) yields the intervened probability, i.e.

$$p(y|do(a)) = \sum_{x_{\mathrm{pa}(a)}} p(y|a, x_{\mathrm{pa}(a)})p(x_{\mathrm{pa}(a)}) \tag{7}$$

this will be proved after defining $p(y|do(a))$ using original distribution $p$.

*Remark.* The existence of effect modifications by some variable(s) $L$ can be identified from causal DAG, but not its magnitude and direction for each level of $L$.

Another way of calculation avoids consulting the mutilated distribution $P_m$ and making assumptions on it, by rigorously defining the mutilated distribution $P_m$ using the original probabilities first. Recall the factorisation of the joint probability of DAG,

$$P(x_V) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)})$$

(note: on a causal diagram, this equation means probability distribution for each variable is only dependent on its direct causes, i.e. parents)
Intervention on treatment $T$ (denote the value as $x_T$) removes parental effects on node $T$, so the term $p(x_t \mid x_{\mathrm{pa}(t)})$ should be removed from the joint distribution.

**Definition 24** (Intervention)**.** The intervened probability distribution (under treatment $T = t$) is

$$\begin{aligned}
p(x_V|do(T = t)) := &\prod_{v \in V \setminus \{T\}} p(x_v \mid x_{\mathrm{pa}(v)}) \quad \text{assuming } x_T = t \\
&= \frac{p(x_v)}{p(t \mid x_{\mathrm{pa}(T)})}
\end{aligned}$$

note if $x_T \neq t$, the probability is set to 0.

*Remark.* If $p(t \mid x_{\mathrm{pa}(T)}) \approx 1$, i.e. $t$ is a natural outcome of its parents' values $x_{\mathrm{pa}(T)}$, then $p(x_V|do(T = t)) \approx p(x_V)$. However, if the intervention $t$ is quite unexpected, the intervened probability is scaled up massively by $1/p(t \mid x_{\mathrm{pa}(T)})$.

The marginal distribution $P(y|do(T=t))$ can be obtained by summing over the parents of node $A$. (equation 7)

**Theorem 5.3** (parent set is valid adjustment)**.** *If $\mathcal{G}$ is causal DAG, then*

$$p(y|do(T=t)) = \sum_{x_{pa(T)}} p(y|t, x_{pa(T)})p(x_{pa(T)})$$

*Proof.* Partition all variables $V$ into $T, Y, \mathrm{pa}(T), W$ (where $W := V \setminus \{A,\, Y\} \cup \mathrm{pa}(T)$). By definition,

$$\begin{aligned}
p(x_V \mid \mathrm{do}(T=t)) = \frac{p(x_V)}{p(t \mid x_{\mathrm{pa}(T)})} &= \frac{p(t, y, x_{\mathrm{pa}(T)}, x_W)}{p(t \mid x_{\mathrm{pa}(T)})} \\
&= \frac{p(y, x_W \mid t, x_{\mathrm{pa}(T)})p(t \mid x_{\mathrm{pa}(T)})p(x_{\mathrm{pa}(T)})}{p(t \mid x_{\mathrm{pa}(T)})} \\
&= p(y, x_W \mid t, x_{\mathrm{pa}(T)})p(x_{\mathrm{pa}(T)})
\end{aligned}$$

By the law of total probability,

$$\begin{aligned}
p(y \mid \mathrm{do}(T=t)) &= \sum_{x_W,\, x_{\mathrm{pa}(T)},\, x_T=t} p(x_V \mid \mathrm{do}(T=t)) \\
&= \sum_{x_W,\, x_{\mathrm{pa}(T)}} p(y, x_W \mid t, x_{\mathrm{pa}(T)})p(x_{\mathrm{pa}(T)}) \\
&= \sum_{x_{\mathrm{pa}(T)}} p(x_{\mathrm{pa}(T)}) \sum_{x_W} p(y, x_W \mid t, x_{\mathrm{pa}(T)}) \\
&= \sum_{x_{\mathrm{pa}(T)}} p(x_{\mathrm{pa}(T)})p(y \mid t, x_{\mathrm{pa}(T)}) \quad \text{by law of total probability again}
\end{aligned}$$

□

## 5.4 Systematic Bias

The term *systematic bias* refers to an additional association between treatment $A$ and target $Y$ that is not from the causal effect. This would make the data insufficient to identify or compute the causal effect even if you have infinite samples.

**Definition 25** (Systematic Bias)**.** There is a systematic bias when

$$P(Y^{a=1}=1) - P(Y^{a=0}=1) \neq P(Y=1|A=1) - P(Y=1|A=0)$$

Lack of exchangeability gives systematic bias, and lack of conditional exchangeability creates conditional bias.

**Definition 26** (Conditional Bias)**.** There is a systematic bias for condition $L$ when

$$P(Y^{a=1}=1 \mid L=l) - P(Y^{a=0}=1 \mid L=l) \neq P(Y=1|A=1, L=l) - P(Y=1|A=0, L=l)$$

There are three main sources of bias

- Confounding: treatment $A$ and outcome $Y$ share a common cause, and the common cause is either unconditioned or unmeasured.
- Selection Bias: when a common effect of $A, Y$ is conditioned
- measurement bias: due to random errors made in measurements

This section will only discuss the first two biases, both of which have been illustrated from the examples in section 5.2.

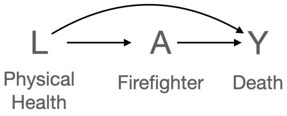


**Figure 35: Direct Confounding**

### 5.4.1 Confounding

Whenever there is a common cause (may not be direct) of treatment $A$ and outcome $Y$, we say there is *confounding*. Below are three typical confounding structures. However, confounding may have more complicated structures.

**Example 13.** In figure 35, the causal effect of being a firefighter $A$ to the risk of death $Y$ is studied. The common cause $L$ means whether the person has good physical health. On the one hand, being a firefighter constantly puts the person in life-threatening danger, so causation $A \to Y$ gives a positive association. But $A$ also has an association with $Y$ through the common cause $L$. (through the open back-door path $A \leftarrow L \to Y$) Physical fitness causes a person to pursue a fire-fighter career but also reduces the risk of death. So firefighters tend to have a lower risk of death from this perspective. The observed data may have a mixture of both associations, and they may cancel out. In that case, the experimenter would mistakenly conclude that $A$ has no causal effect on $Y$.

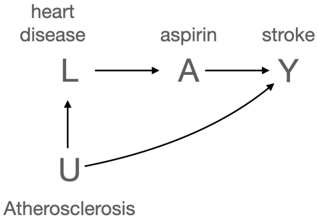


**Figure 36: Confounding by channelling/indication**

**Example 14.** Atherosclerosis is the thickening or hardening of arteries that potentially cause both heart disease (restricts blood supply to the heart) and stroke(restricts blood supply to the brain). Aspirin is a medicine used to reduce the risk of heart attacks, commonly used by people with heart disease. The causal relations are summarised in figure 36. Note the variable $U$ (atherosclerosis) is unmeasured. Checking atherosclerosis requires careful examination of arteries, which is often impossible. But $U$ causes $A$ (through the mediator $L$) and causes $Y$ directly, i.e. $U$ is a common cause, so there is confounding. Association flows through the causation $A \to Y$ and the back-door path $A \leftarrow L \leftarrow U \to Y$. The association through the back-door path distracts our identification of a causal effect. So there is a bias.

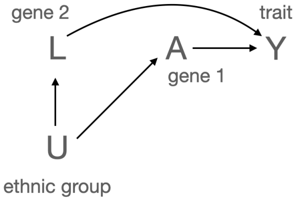


**Figure 37: Linkage Disequilibrium/population stratification**

**Example 15.** In figure 37, the effect of a gene haplotype $A$ on a trait $Y$ (for example, $Y$ could be the colour of your hair) is studied. But there is also another gene ($L$) causing trait $Y$. The ethnic group $U$ of individuals affects whether the two-gene haplotype occurs. Association flows through the back-door path $A \leftarrow U \to L \to Y$, because the two gene haplotypes may rise together more frequently in an ethnic group, or both missing in another ethnic group. In this example, $U$ directly causes $A$ and causes $Y$ through mediator $L$. So $U$ still counts as a common cause and there is confounding. Such structure is called *linkage disequilibrium* or *population stratification*.

To summarise, the existence of an open back-door path brings bias to the causal path. The rigorous definitions are given below.

**Definition 27** (causal path)**.** A causal path from $A$ to $Y$ is a path on the causal DAG where all arrows point in the same direction. So the path looks like $A \to \cdots \to Y$. All nodes on this path other than treatment $A$ are called *causal node.* The set of causal nodes on all causal paths on $\mathcal{G}$ is denoted $\mathrm{cn}_{\mathcal{G}}(A \to Y)$.

**Definition 28** (back-door path)**.** A back-door path from $A$ to $Y$ is any path between $A, Y$ that is not causal. i.e. some arrows point in the opposite direction.

Marginal exchangeability $Y^a \perp A$ does not hold if an open back-door path exists. But we have seen that in simple cases like figure 30, conditioning on the direct common cause $L$ blocks the association flow and removes the bias. In general, is it possible to find a set $C \subseteq V \setminus \{A, Y\}$ such that conditional exchangeability $Y^a \perp A \mid L$ holds? Such $L$ allows us to find the causal effect by adjustment on $L$. Pearl proposed the back-door criterion (BDC), which specifies when the back-door path is blocked.

**Definition 29** (Back-door criterion)**.** For causal DAG, $C \subseteq V \setminus \{A, Y\}$ satisfies the back-door criteria(BDC) if

- all back-door paths are blocked after conditioning on $C$
- $C$ has no descendant of $A$

The second criterion prevents the following situation: if $L$ is a common effect of $A, Y$, the back-door path is $A \to L \leftarrow Y$. But it is blocked by collider $L$. If we condition on $L$, a descendant of $A$, it opens this path and introduces a bias. So no descendant of treatment $A$ is allowed in the set $C$.

**Theorem 5.4** (Back-door criterion and conditional exchangeability)**.** *Assuming the causal DAG satisfies the Markov property. If $C$ satisfies BDC, then conditional exchangeability holds given $C$. (i.e. causal effect can be identified by conditioning on $C$)*

*Remark.* Under FFRCISTG(Rubin's model), the converse also holds. i.e. whenever conditional exchangeability holds, BDC is satisfied. However, in Pearl's model NPSEM-IE, this does not hold due to the assumption of independent errors.

The theorem can be proved easily with the help of SWIG (see section 5.4.2).

BDC is satisfied when

- There is no confounding
- No unmeasured confounding: $C$ suffices to block all back-door paths, such $C$ is called *valid set for confounding adjustment*

and when BDC is satisfied, $C$ can be used for adjustment (using standardisation). Therefore, $C$ satisfying the BDC is also called *back-door adjustment set.*

**Example 16.** In figure 35, $C = \{L\}$ can be used for adjustment. In figure 36, $C = \{L\}$ can also be used for adjustment, even though $L$ is not the common cause of $A$ and $Y$. So the probability after intervention is

$$p(y|do(a)) = \sum_{x_L} p(x_L) p(y|a, x_L)$$

In general, a *valid adjustment set* is a set of variables $C$ for which conditioning on it and summing over all possible values of $C$ (standardisation by $C$) recovers the causal effect.

**Definition 30** (valid adjustment set)**.** $C$ is valid adjustment set for ordered pair $(a, y)$ if

$$p(y|do(a)) = \sum_{x_C} p(x_C) p(y|a, x_C)$$

There are other adjustment methods, mainly in two categories

- G-methods: standardisation, IP weighting, G-estimation
- Stratification: estimate effect of $A$ on $Y$ for different levels of $L$. (and not bother about the overall population effect)

$L$ is said to be a confounder (or confounder set) if $\{A, Y, L\}$ is identifiable but

$$\{A, Y\}$$

is not identifiable. So a confounder does not need to be the common cause of $A, Y$, we only care about variables that are sufficient for adjustment.

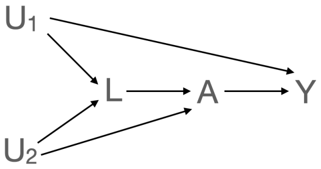


**Figure 38: Confounding with two backdoor paths**

**Example 17.** We study an interesting example where no valid adjustment set exists. In figure 38, $U_1$ is a common cause of $L$ and $Y$, $U_2$ is a common cause of $A$ and $L$. We assume $U_1$, $U_2$ are unmeasured. Suppose $C$ is a valid adjustment set, you may propose $L \in C$. Indeed, conditioning on $L$ blocks the back-door path $A \leftarrow L \leftarrow U_1 \rightarrow Y$. But it opens the backdoor path $A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$ which was blocked by $L$ with no conditioning. But if $L \notin C$ (not conditioning on $L$) the back-door path $A \leftarrow L \leftarrow U_1 \rightarrow Y$ if left open. So no valid adjustment set exists.

There is a way to solve it, but one needs to find measurable indicator $L_1$ between $U_1, L$ or $U_1, Y$. i.e. $U_1 \rightarrow L_1 \rightarrow L$ or $U_1 \rightarrow L_1 \rightarrow U$. Then $C = \{L_1, L\}$ is a valid adjustment set. (think of why) Similarly, for any indicator $L_2$ between $U_2, A$ or $U_2, L$, $C = \{L, L_2\}$ is a valid adjustment set. (see figure 39)

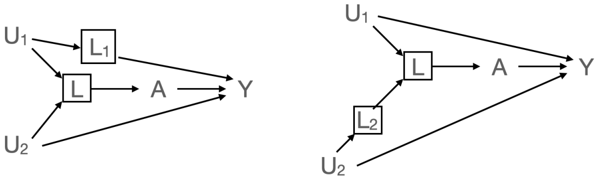


**Figure 39: Resolution to confounding with two backdoor paths**

If the confounder is unknown or unmeasured, the effect of confounding can be roughly estimated by sensitivity analysis [24]. Or if there is a measured descendant of the confounder, conditioning on it lightens part of the bias.

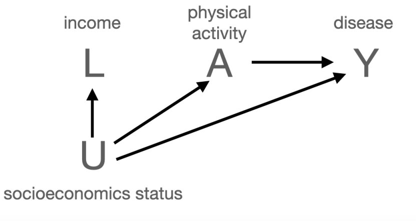


**Figure 40: Surrogate confounder**

**Example 18.** Figure 40 studies the causal effect of physical activity on the probability of getting a disease. However, the correlation is disrupted by socioeconomic status $U$ (position in the society regarding ability to access the resources), which is called *causal confounder*. However, it is tough to measure $U$ directly. But $U$ usually decides the person's income $L$, which is easy to measure. Conditioning on $L$ does not eliminate bias introduced by $U$ completely, but income does provide some information on socioeconomic status $U$. Such $L$ is called *surrogate confounder* (a confounder that is not a cause of both the treatment and outcome).

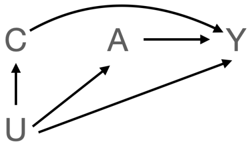


**Figure 41: Pre-outcome $C$ and equi-confounding assumption**

**Confounding Adjustment by Equi-confounding Assumption**

Adjustment of confounding becomes easy under the equi-confounding assumption formulated as below (figure 41). Suppose a pre-outcome $C$ is measured ($C$ does not cause treatment $A$), due to confounding by $U$, $C$ and $A$ are not independent. If the magnitude of confounding (by $U$) is the same on $A, C$ and $A, Y^0$, i.e.

$$E(Y^0 \mid A = 1) - E(Y^0 \mid A = 0) = E(C \mid A = 1) - E(C \mid A = 0)$$

we say there is an equi-confounding effect. Under this assumption, the following holds

$$\begin{aligned} E(Y^1 - Y^0 \mid A = 1) &= E(Y - C \mid A = 1) - E(Y - C \mid A = 0) \\ &= \underbrace{(E(Y \mid A = 1) - E(Y \mid A = 1))}_{\text{observed association}} - \underbrace{(E(C \mid A = 1) - E(C \mid A = 0))}_{\text{confounding effect}} \end{aligned}$$

i.e. we can remove confounding effects using a pre-outcome $C$ which also suffers from confounding up to the same extent. The formula is called *difference-in–difference*, as illustrated in Figure 42. The bottom line is the untreated group, and the top line is the treated group, where the gradient of the line changes after treatment. A more detailed introduction can be found in [1].

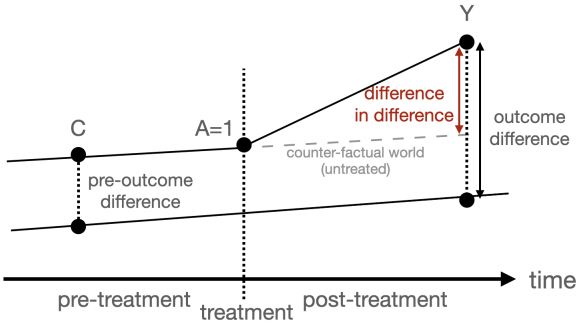


**Figure 42: Illustration of Difference in difference method (figure inspired by [1])**

More surprisingly, if one has a negative outcome control $C$ (pre-outcome) and negative treatment control $Z$ (pre-treatment), the causal effect can be identified even with the unmeasured confounder $U$ (with some additional conditions).(see figure 43)The identification process is called *proximal causal inference* [2], a relatively new branch of causal inference under active research.

### 5.4.2 Single World Intervention Graph

[17] introduces Single World Intervention Graph(SWIG), which incorporates counterfactuals into the causal graphs. It represents the counterfactual world created by a single intervention. An example of SWIG is given in Figure 44.

The treatment node $A$ is split into two (un-connected) nodes:

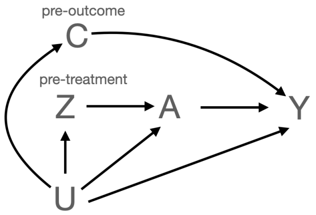


**Figure 43: Identification of causal effect with pre-treatment and a measured pre-outcome**

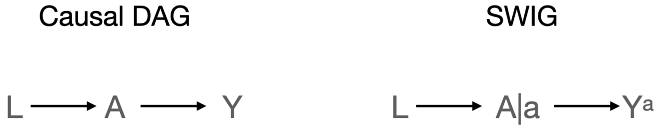


**Figure 44: The SWIG in a simple case**

- $A$: inherits all edges into $A$ on the causal DAG.
- $a$: inherits all edges out of $A$ in causal DAG, and all descendants of $a$ becomes the counterfactual version under $a$.

The stick between $A$ and $a$ represents they were the same node in the causal DAG. But $A$ and $a$ are treated as **non-adjacent** nodes on the SWIG because the focus of SWIG is to study bias (on the back-door paths for example). So the causal path is not under consideration, and it is cut off by the stick.
Note $a$ is a constant node and the edge $a \to Y^a$ does not represent a causation. Non-descendants of $A$ are not affected (under faithfulness assumption) by intervention $A = a$.

**Lemma 5.5.** *Edges except the causal paths are all preserved in the SWIG, so any statistical independence along the back-door paths remains the same.*

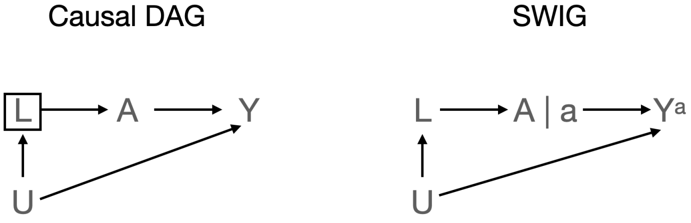


**Figure 45: The SWIG for confounding**

**Example 19.** Suppose there is a confounding, as shown in the left of figure 45. Recall that conditioning on $L$ adjusts the confounding effect. On the SWIG (right of figure 45), $Y^a$ and $A$ are d-separated by $L$, so (assuming Markov property), $Y^a \perp A \,|\, L$. i.e. conditional exchangeability holds.

As shown in the example above, the Back-door criterion becomes ”no open path between $A$ and $Y$” on the SWIG, equivalent to $A$, $Y^a$ are d-separated $\Leftrightarrow$ conditional exchangeability (under faithfulness). So by lemma 5.5, theorem 5.4 (BDC implies valid adjustment) is trivially true.

**Example 20.** We examine a case where conditioning on the descendant of treatment $A$ does not yield conditional exchangeability. There is confounding in Figure 46, and $L$ is a mediator of $A$, $Y$ as well. Conditioning on $L$ does

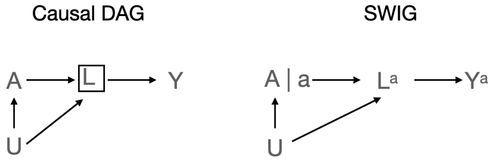


**Figure 46: confounding case with mediator $M$ and its SWIG**

block the back-door path, but it also disrupts the causal path $A \to L \to Y$. From the SWIG of this case, we see $Y^a \perp A \,|\, L^a$, but this is NOT equivalent to $Y^a \perp A \,|\, L$. Conditional exchangeability cannot be concluded in this case, and the systematic bias persists.

Later, we will introduce the so-called *forbidden set*, which is a set of nodes that we should not condition when adjusting for systematic bias.

In addition to the back-door criterion, Pearl also proposed a front-door formula to deal with the case where mediators exist on the causal path (from $A$ to $Y$). Adjustment by standardisation does not work as we have seen in the example above. The proof can be done using SWIG. Note there are two possible interventions, on $M$ and $A$ respectively. So two SWIGs are involved.

**Theorem 5.6** (front-door adjustment)**.** *Consider the causal DAG in Figure 46. Suppose $Y^a$, $M^a\, Y^m$ are well-defined, we have the front-door formula*

$$P(Y^a = 1) = \sum_m P(M = m|A = a) \sum_{a'} P(Y = 1 \,|\, M = m, A = a')P(A = a')$$

*It is similar to standardisation, with an extra layer on node $A$. The formula adjusts along the causal path (front-door path) $A \to M \to Y$, aligning with its name.*

*Proof.*

$$\begin{aligned} P(Y^a = 1) &= \sum_m P(M^a = m)P(Y^a = 1 \,|\, M^a = m) \quad \text{by law of total probability} \\ &= \sum_m P(M = m \,|\, A = a)P(Y^a = 1 \,|\, M^a = m) \quad \text{no confounding between } A, M \end{aligned}$$

Further, if $M^a = m$, then $Y^a = Y^m$ because $M$ is the direct cause of $Y$. From the SWIG to the left of figure 47, $Y^m = Y^a$ and $M^a$ are completely separated, so $Y^m \perp M^a$.

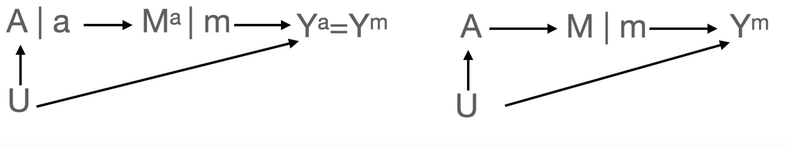


**Figure 47: Two SWIGs corresponding to interventions on $A$(left) and $M$(right) respectively**

So we can simplify

$$\begin{aligned} P(Y^a = 1 \,|\, M^a = m) &= P(Y^a = 1) = P(Y^m = 1) \\ &= \sum_{a'} P(Y^m = 1 \,|\, A = a')P(A = a') \end{aligned}$$

From the SWIG to the right of figure 47, $Y^m$ and $M$ are d-separated by $A$, so $Y^m \perp M \mid A$.

$$\begin{aligned} P(Y^a = 1 \mid M^a = m) &= \sum_{a'} P(Y^m = 1 \mid M = m, A = a')P(A = a') \\ &= \sum_{a'} P(Y = 1 \mid M = m, A = a')P(A = a') \quad \text{by consistency of } Y \end{aligned}$$

□

### 5.4.3 Selection Bias

Selection bias happens when the collected data is a set that brings association to $A$ and $Y$. Figure 48 shows a case where $A$, $Y$ are not associated, but the data set you collected happens to capture those with consistently low or high values of $A, Y$. (e.g. maybe other individuals are unwilling to participate) So your study concludes that $A$, $Y$ are positively correlated, which is wrong!

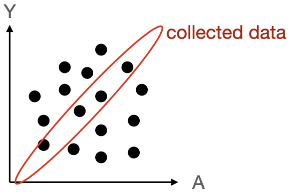


**Figure 48: Selection bias that brings positive correlation**

Now we study some common structures of selection bias (on causal DAGs)

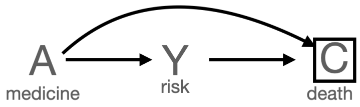


**Figure 49: Selection bias by common effect**

**Example 21.** A medicine $A$ reduces the risk of death $C$ by reducing the risk $Y$, but it also reduces the risk of death by reducing other risks. (figure 49) For some reason, only data of those alive ($C = 0$) are collected. Patients with higher risks ($Y = 1$) who survived are more likely to have taken the medicine. So a positive correlation(selection bias) is introduced between $A$, $Y$. We say conditioning on $C$ opens a back-door path.

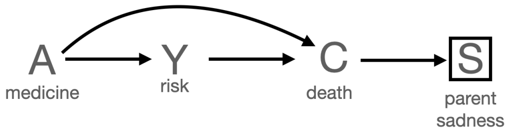


**Figure 50: Selection bias by descendant of common effect**

**Example 22.** Consider the same situation as in 49, but this time researchers attempt to collect data for patients who have passed away from their parents. $S = 1$ represents the parents who are sad and unwilling to provide data. So only $S = 0$ is selected for analysis. (figure 50) Again, a positive correlation(selection bias) is introduced between $A$, $Y$.

*Remark.* In the above two examples, the bias exists even if $A$ does not cause $Y$. The same applies to confounding. So confounding and selection bias are called *bias under null*.

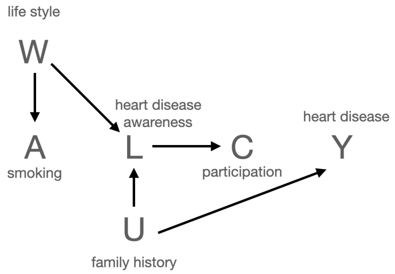


**Figure 51: Selection bias by a descendant of a collider on the back-door path**

**Example 23.** 51 represents a heart-disease survey study conducted. Awareness of heart disease ($L = 1$) causes people to participate ($C = 1$). The awareness is caused by lifestyle $W$ as well as the family history of heart disease $U$ (unmeasured). A bad lifestyle causes someone to smoke ($A$) and a family history of heart disease causes a higher possibility of getting heart disease $Y$. See the causal DAG in figure 51. No matter what causal relation between $A$, $Y$ is, there is a back-door path $A \leftarrow W \rightarrow L \leftarrow U \rightarrow Y$ opened because the data is only collected for those who participated (conditioning on $C = 1$), which is a descendant of the collider $L$. Here, neither $L$ nor $C$ is a common effect of $A$ or $Y$, but selection bias is introduced.

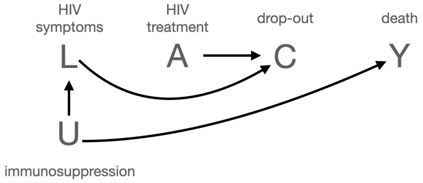


**Figure 52: Selection bias by a descendant of treatment**

**Example 24.** In contrast to the above example, 52 is where selection bias is introduced by drop-out $C$, an effect of HIV treatment $A$. Because side effects of treatment cause some patients to quit the experiment ($C = 1$). Unmeasured quantity $U$ is immuno-suppression which causes some symptoms of HIV $L$ and death $Y$. People with bad symptoms $L$ tend to drop out ($C = 1$). See the causal DAG in figure 52. Data are only collected for patients who do not drop out, i.e. conditioning on $C = 0$. But $C$ is a collider on the back door path $A \rightarrow C \leftarrow L \leftarrow U \rightarrow Y$. So selection bias is introduced.

.

Selection bias even exists for randomised experiments, because the randomisation is a pre-treatment process, but selection bias are usually introduced by post-treatment factors.

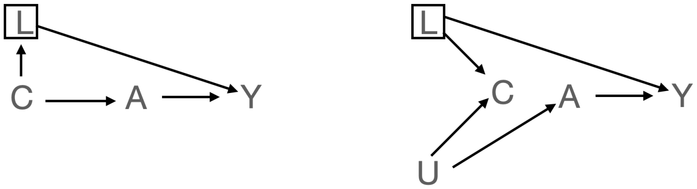


**Figure 53: the same adjustment may work for both confounding and selection bias**

Selection bias and confounding are both systematic biases, and they may have the same degree of effect. For the selection bias in the right of figure 53, conditioning on $L$ blocks the back-door path. The same treatment works for the confounding case on the left of figure 53.

However, there are advantages to distinguishing these two biases.

- adjustments could be different in some cases for selection bias and confounding. (incorrect adjustments, e.g. by stratification, may introduce new bias)
- this guides study design, in terms of better selection of samples
- explains why a pre-treatment sometimes behaves like a confounder. (a pre-treatment may affect the willingness to participate, which may introduce a selection bias)

**Adjustments for Selection Bias**
As shown in figure 53, if there is any measured node on the back-door path with collider $C$ that is not a collider ($L$ in this case), then conditioning on it removes the selection bias. So this is an adjustment by stratification, and the population effect can be found by standardisation (summing over levels of $L$)

However, in the case 54 where selection bias is introduced by $C$, the only possible node to stratify is $L$. But $L$ is a collider on the back-door path. So stratification does not adjust the selection bias.

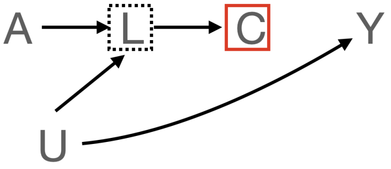


**Figure 54: Selection bias not removable by stratification. The red box is selection bias, dotted box is our attempt to stratify**

In such cases, IP weighting is required. It adjusts by reweighing according to treatment and values of $L$, making the probability unconditional (see Chapter 2.4 of [7])

## 5.5 Adjustment Criterions

Recall a valid adjustment set $C$ is a set on which conditioning and summing over levels of $C$ identifies the causal effect from observational data. i.e.

$$p(y|do(a)) = \sum_{x_C} p(x_C)p(y|a, x_C)$$

**Lemma 5.7.** *Parent set $pa(T)$ is a valid adjustment set for $T \to Y$*

*Proof.* the same as theorem 5.3. □

Pearls proposed the back-door criterion(BDC) (all back-door paths blocked, and no descendant of the treatment $A$ conditioned) for valid adjustment. If conditional exchangeability (conditioning on $C$) holds, then $C$ is a valid adjustment set. So by Theorem 5.4, any set $C$ satisfying BDC is a valid adjustment set. We say this criterion is *sound*. An elementary proof of this statement without using conditional exchangeability is given below.

**Theorem 5.8** (Back-door set is valid adjustment set)**.** *Suppose a set $C$ satisfies BDC, then $C$ is a valid adjustment set.*

*Proof.* Parent set $\mathrm{pa}(T)$ is a valid adjustment set, so the proof is complete if we can show adjustment using $x_C$ is the same as adjustment using $x_{\mathrm{pa}(T)}$, i.e.

$$\sum_{x_{\mathrm{pa}(T)}} p(x_{\mathrm{pa}(T)})p(y \mid t, x_{\mathrm{pa}(T)}) = \sum_{x_C} p(x_C)p(y \mid t, x_C) \qquad \text{(Target equation)}$$

Using the law of total probability we can put $x_C$ into LHS

$$\begin{aligned}\sum_{x_{\mathrm{pa}(T)}} p(x_{\mathrm{pa}(T)})p(y\,|\,t,x_{\mathrm{pa}(T)}) &= \sum_{x_{\mathrm{pa}(T)}} p(x_{\mathrm{pa}(T)})\sum p(y,x_C\,|\,t,x_{\mathrm{pa}(T)}) \\ &= \sum_{x_{\mathrm{pa}(T)}} p(x_{\mathrm{pa}(T)})\sum p(y\,|\,x_C,t,x_{\mathrm{pa}(T)})p(x_C\,|\,t,x_{\mathrm{pa}(T)}) \qquad (1)\end{aligned}$$

Similarly, we can put $x_{\mathrm{pa}(T)}$ to RHS of the target equation

$$\begin{aligned}\sum_{x_C} p(x_C)p(y\,|\,t,x_C) &= \sum_{x_C} p(y\,|\,t,x_C)\sum_{x_{\mathrm{pa}(T)}} p(x_C) \\ &= \sum_{x_C} p(y\,|\,t,x_C)\sum_{x_{\mathrm{pa}(T)}} p(x_C\,|\,x_{\mathrm{pa}(T)})p(x_{\mathrm{pa}(T)}) \quad \text{then exchange the summation,} \\ &= \sum_{x_{\mathrm{pa}(T)}} p(x_{\mathrm{pa}(T)})\sum_{x_C} p(y\,|\,t,x_C)p(x_C\,|\,x_{\mathrm{pa}(T)}) \qquad (2)\end{aligned}$$

Comparing equations (1) and (2), our goals are proving

$$p(y\,|\,x_C,t,x_{\mathrm{pa}(T)}) = p(y\,|\,t,x_C), \quad \text{and } p(x_C\,|\,t,x_{\mathrm{pa}(T)}) = p(x_C\,|\,x_{\mathrm{pa}(T)}) \quad (*)$$

Assuming $C \cap x_{\mathrm{pa}(T)} = \emptyset$ (adjustment set does not contain any parent of $T$), then proving $Y \perp \mathrm{pa}(T)\,|\,C,T$ and $T \perp C\,|\,\mathrm{pa}(T)$ is enough.

All nodes of $C$ are non-descendant of $T$ (by the second condition of back-door criterion), i.e. $C \subseteq \mathrm{nd}_{\mathcal{G}}(T)$. By local Markov property, $T \perp \mathrm{nd}_{\mathcal{G}}(T)\,|\,\mathrm{pa}(T)$, so $T \perp C\,|\,\mathrm{pa}(T)$ is true.

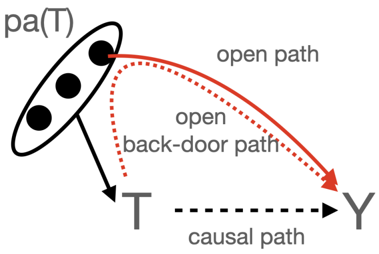


**Figure 55: Illustration of proving $Y \perp \mathrm{pa}(T)\,|\,C,T$ by contradiction. The dotted line means we ignore the structure of the path in between.**

In terms of d-separation, another goal $Y \perp \mathrm{pa}(T)\,|\,C,T$ is equivalent to: $Y$ is d-separated from $\mathrm{pa}(T)$ by condition set $C$ and treatment $T$. Suppose for contradiction that there is an open path $\pi$ from $Q \in \mathrm{pa}(T)$ to $Y$ not going through $C \cup \{T\}$. The causal path is not in $\pi$ because $T \notin \pi$. So we have an open back-door path from $T$ to $Q$, then through path $\pi$ to $Y$. (see Figure 55) By the back-door criterion, $C$ blocks all open back-door paths. So at least one node of $C$ is on the path $\pi$. Contradiction.

Suppose $C$ contains parent(s) of $T$, i.e. $C \cap x_{\mathrm{pa}(T)} \neq \emptyset$, then prove $Y \perp \mathrm{pa}(T) \setminus \{C\}\,|\,C,T$ instead also implies our goals (*). The proof is similar to the above (left as an exercise). Another independence $T \perp C\,|\,\mathrm{pa}(T)$ still holds for the same reason. □

But the back-door criterion is not *complete* in the sense that some valid adjustment sets do not satisfy BDC.

**Example 25.** In figure 56, $\{L\}$ is a valid adjustment set (try to prove it!), but it does not satisfy the back-door criterion because $L$ is a (direct) descendant of $A$.

In contrast, the generalised adjustment criterion[21] is complete.

It starts by defining some nodes that cannot be used for adjustment (by standardisation). These are the nodes that we should NOT condition on.

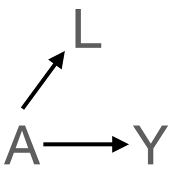


**Figure 56: An example where valid adjustment set $\{L\}$ does not satisfy BDC**

- Any node on the causal path from $A$ to $Y$ (mediators)
- Any consequence of the outcome (post-outcome), may introduce selection bias or break exchangeability (see example 20)
- Any consequence of the mediators (nodes on the causal path from $A$ to $Y$ except $A$ and $Y$), for the same reason as above. (draw a SWIG if you are not sure)
- Treatment $A$ and outcome $Y$

**Definition 31** (Forbidden nodes)**.** For causal DAG $\mathcal{G}$, the set of forbidden nodes for the study of causal relation $A \to Y$ is

$$\mathrm{forb}_{\mathcal{G}}(A \to Y) := \mathrm{de}(\mathrm{cn}_{\mathcal{G}}(A \to Y)) \cup \{A\}$$

where the set $\mathrm{cn}_{\mathcal{G}}(A \to Y)$ is the set of causal nodes. (nodes on the causal path from $A$ to $Y$ other than $A$ itself)

*Remark.* Any causal node (in $\mathrm{cn}_{\mathcal{G}}(A \to Y)$) is a descendant of $A$, so any forbidden node is a descendant of $A$, i.e.$\mathrm{forb}_{\mathcal{G}}(A \to Y) \subseteq \mathrm{de}_{\mathcal{G}}(T)$.

Using the forbidden set, the *generalised adjustment criterion* can be defined

**Definition 32** (generalised adjustment criterion)**.** A set $C$ satisfies the generalised adjustment criterion(GAC) w.r.t. $(T, Y)$ if

- $C$ does not contain any node in the forbidden set $\mathrm{forb}_{\mathcal{G}}(T \to Y)$
- every back-door(non-causal) path from $T$ to $Y$ is blocked by $C$

*Remark.* If we change the forbidden set to $\mathrm{de}_{\mathcal{G}}(T)$, the back-door criterion is obtained. By the last remark, any set $C$ satisfying BDC also satisfies GAC, so GAC is indeed a generalisation of BDC.

**Example 26.** The parent set $\mathrm{pa}_{\mathcal{G}}(T)$ satisfies GAC. Clearly, it does not have any descendant of $T$, so no forbidden node. Suppose there is a back-door path(non-causal path) not going through $\mathrm{pa}_{\mathcal{G}}(T)$ (Figure 57), then the first edge is going out of $T$, say to $S$. There must eventually be a collider on the way from $S$ to $Y$ ($S$ can also be a collider), otherwise, the path is causal. These back-door paths are naturally blocked by the collider, other back-door paths (going through $\mathrm{pa}_{\mathcal{G}}(T)$) are blocked by conditioning on $\mathrm{pa}_{\mathcal{G}}(T)$. Therefore, $\mathrm{pa}_{\mathcal{G}}(T)$ blocks all open back-door paths. i.e. $\mathrm{pa}_{\mathcal{G}}(T)$ satisfies BDC and GAC.

In the rest of this section, we will prove GAC is sound and complete.

**Theorem 5.9** (soundness)**.** *If on DAG $\mathcal{G}$, $C$ satisfies the generalised adjustment criterion w.r.t. $(T, Y)$. Then $C$ is a valid adjustment for the causal relation $T \to Y$.*

We prove this by reducing it to the case of BDC. First, we show any descendant of $T$ in $C$ is not required to block the open back-door paths. i.e. the set $B = C \cap \mathrm{nd}_{\mathcal{G}}(T)$ also satsifes GAC (Hence, $B$ is a back-door set). Then we show adjustment by $B$ is the same as an adjustment by $C$.

*Proof.* .

**Step 1.** $B = C \cap \mathbf{nd}_{\mathcal{G}}(T)$ **is a back-door set**

For any node $b \in B$, it is a descendant of $T$ not on the causal path. Consider paths from $d$ to $Y$ not involving $T$ or any ancester in $\mathrm{de}_{\mathcal{G}}(T)$. There are only three types of relations between $d$ and $Y$:

- no path from $d$ to $Y$. So $d$ is not on any back-door path

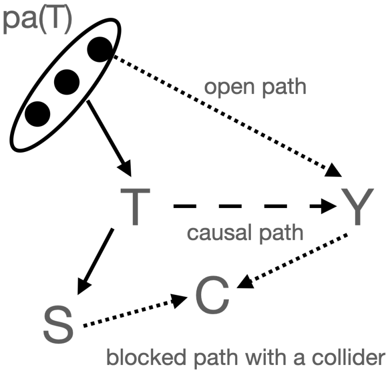


**Figure 57: Possible back-door paths categorised by whether it contains $\mathrm{pa}_{\mathcal{G}}(T)$ or not. (dotted lines are paths with any structure between allowed, they may be additional nodes in-between)**

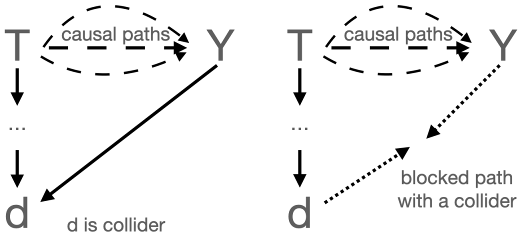


**Figure 58: Descendants of $T$ are not on any open back-door path (again, dotted lines are paths which we do not care about the structure in between)**

- $Y \to d$, $d$ is a collider (left of figure 58). But this is impossible, conditioning on such $d$ opens a back-door path, contradicting that $C$ satisfies GAC.
- There is at least one path from $d$ to $Y$ (right of figure 58). But since $d \in C$, it cannot be on a causal path. So all these paths from $d$ to $Y$ are blocked by at least one collider.

Therefore, any back-door path from $T$ to $Y$ through $d$ is naturally blocked. Removing descendants of $T$ from $C$ will not affect its ability to block open back-door paths. So conditioning on $B = C \cap \mathrm{nd}_{\mathcal{G}}(T)$, there is no open back-door path. Since $B$ does not contain descendants of $T$, it satisfies BDC.

**Step 2. Adjustment by $B$ and $C$ are the same**

The aim is to prove

$$\sum_B p(x_B)p(y \mid t, x_B) = \sum_C p(x_C)p(y \mid t, x_C) \qquad \text{(target equation)}$$

This can be done by removing $D := C \setminus B$ (descendants of $T$ in $C$) node by node.

First, pick $d$ deepest down in the set $D$ in the sense that $\mathrm{de}(d) \cap C = \emptyset$. (see figure 59) As discussed before, there is either no path from $d$ to $Y$ or a path blocked by a collider. So open path from $d$ to $Y$ either goes through $T$ or a node in $\mathrm{de}(d)$ above it that is not in $D$. (for example in figure 59, if there is an open path from $Y$ to $Z \in \mathrm{de}(d)$, then this open path continues down to $d$) The first case corresponds to $d \perp_d Y \mid t, C \setminus \{d\}$ ($\perp_d$ means d-separation, which implies independence if the model is Markov on $\mathcal{G}$) and the second corresponds to $d \perp_d T \mid C \setminus \{d\}$.

Case 1. $d \perp_d Y \mid t, C \setminus \{d\}$

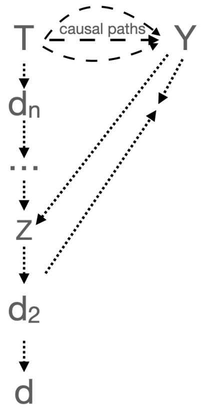


**Figure 59: Removal of descendants of $T$**

$$\begin{aligned}\sum_{x_C} p(x_C)p(y \mid t, x_C) &= \sum_{x_{C\setminus d}, x_d} p(x_{C\setminus d}, x_d)p(y \mid t, x_{C\setminus d}, x_d)\\ &= \sum_{x_{C\setminus d}, x_d} p(x_{C\setminus d}, x_d)p(y \mid t, x_{C\setminus d}) \quad \text{by independence}\\ &= \sum_{x_{C\setminus d}} p(x_{C\setminus d})p(y \mid t, x_{C\setminus d}) \quad \text{law of total probability}\end{aligned}$$

Case 2. $d \perp_d T \mid C \setminus \{d\}$

$$\begin{aligned}\sum_{x_C} p(x_C)p(y \mid t, x_C) &= \sum_{x_{C\setminus d}, x_d} p(x_{C\setminus d})p(x_d \mid x_{C\setminus d})p(y \mid t, x_C)\\ &= \sum_{x_{C\setminus d}, x_d} p(x_{C\setminus d})p(x_d \mid x_{C\setminus d}, t)p(y \mid x_d, x_{C\setminus d}, t) \quad \text{by independence}\\ &= \sum_{x_{C\setminus d}} p(x_{C\setminus d}) \sum_{x_d} p(x_d \mid x_{C\setminus d}, t)p(y \mid x_d, x_{C\setminus d}, t)\\ &= \sum_{x_{C\setminus d}} p(x_{C\setminus d}) \sum_{x_d} p(y, x_d \mid x_{C\setminus d}, t)\\ &= \sum_{x_{C\setminus d}} p(x_{C\setminus d})p(y \mid x_{C\setminus d}, t) \quad \text{law of total probability}\end{aligned}$$

After removing $d$, the next one ($d_2$ in figure 59) becomes a node with no descendant in $C \setminus \{d\}$. So similar procedure applies until all nodes in $D$ are removed from $C$.

**Step 3. Summary**

Since the back-door set $B$ in $C$ satisfies BDC (by step 1), by theorem 5.8,

$$p(y \mid \mathrm{do}(T = t)) = \sum_B p(x_B)p(y \mid t, x_B) = \sum_C p(x_C)p(y \mid t, x_C)$$

where second equality is by step 2. □

**Definition 33** (Induced Graph)**.** A DAG $\mathcal{G}$ is the induced graph of a causal model(NPSEM) with probability distribution $p$ if $p$ is Markov w.r.t $\mathcal{G}$.

*Remark.* Every causal model induces a DAG (see section 5.2 Causal Diagrams and Pearl's model). However, a DAG can be induced by several possible causal models.

In fact, for every causal model $M$ inducing $\mathcal{G}$, set $C$ satisfying GAC is a valid adjustment set under $M$. (see proof in [21] that uses the twin network)

**Theorem 5.10** (Completeness)**.** *Suppose DAG $\mathcal{G}$ has a set $C$ not satisfying GAC for the pair $(T, Y)$, then there is a causal model $M$ inducing $\mathcal{G}$ where $C$ is not a valid adjustment set.*

The proof proceeds by constructing a probability distribution that violates adjustment formulae. See detailed proof in [21].

**Corollary 2.** *Any valid adjustment set does not contain forbidden nodes.*

## 5.6 Parametric Models

All previous sections are focusing on estimating the causal effect from the conditional probabilities. In the frequentist's view, these probabilities are frequencies in the population. For example, $P(Y = 1 \mid A = 0)$ is the frequency of $Y = 1$ values in the untreated population (divided by the population size). But in reality, the data from the whole population is never available. Researchers take a small sample of data and hope it represents the whole population. In such cases, we obtain an estimated conditional probability $\hat{P}(Y \mid A)$ (or more generally, estimate the conditional expectation $E(Y \mid A)$), usually defined by the sample average. If enough samples are collected, a frequentist's confidence interval can be constructed for this probability.

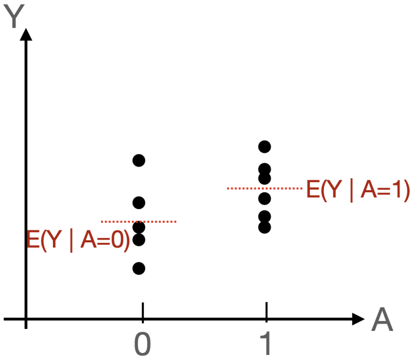


**Figure 60: A non-parametric model for dichotomous $A$ and continuous outcome $Y$. the red dotted lines are the estimates of two conditional expectations**

When the treatment $A$ is discrete and each level of $A$ has enough samples, estimating the sample average for each level of $A$ is quite reasonable. If $A$ is dichotomous(two potential levels), two estimates $\hat{E}(Y \mid A = 0), \hat{E}(Y \mid A = 1)$ are formed. (see Figure 60) The model is called *non-parametric* because no parameter is explicitly assigned. Strictly speaking, each level of $A$ is assigned with a parameter $\hat{E}(Y \mid A = a)$ estimating the conditional expectation $E(Y \mid A = a)$, so the model is *saturated* in terms of having the same number of parameters as the number of conditional means to be estimated.

However, if the treatment $A$ is continuous, for example, the dose of treatment in mg, is such estimation still sensible? Technically, there are infinite levels of possible treatments and infinite conditional means to estimate. To explain this, if the data collected have $A \in \{10, 20, 30, 40, 50\}$. The conditional expectation $E(Y \mid A = 15)$ cannot be evaluated because no data have $A = 15$. So we need a restriction on the distribution of $E(Y \mid A)$ across different values of $A$ to borrow information from other points.

### 5.6.1 Linear Models

The simplest choice for modelling $E(Y \mid A)$ is the linear model (figure 61)

$$E(Y \mid A) = \beta_0 + \beta_1 A \tag{8}$$

, encoding the restriction that $E(Y \mid A)$ must increase linearly with $A$. The coefficients $\beta_0$ and $\beta_1$, called *parameters*, can be estimated from the data by minimising the sum of squared residuals. (residual of each data point is the

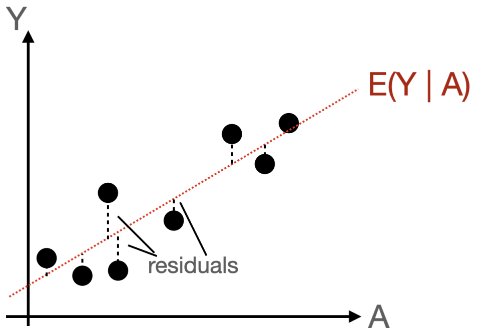


**Figure 61: the parametric model on the conditional expectation and the dotted red line is the predicted line. The residuals are shown as broken black lines.**

distance from the data point to the model predicted value $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 A$, the estimated parameters $\hat{\beta}_0, \hat{\beta}_1$ are chosen to estimate the sum of the square of residuals) This is the *least square method*.

Model (8) is called a *parametric model* because it uses parameters to impose restrictions on the distribution of $E(Y|A)$. The *non-parametric model* we have been using uses no parameter and allows the value of $Y$ to fly everywhere. Of course, the observed $Y$ will not be exactly on the predicted line due to random errors. Recall when adjusting for covariate $L$, an estimation of $E(Y|A, L)$ is required. A parametric linear model on this could be

$$E(Y \mid A, L) = \beta_0 + \beta_{ay\cdot l} A + \beta_{al\cdot y} L$$

where $\beta_{ay\cdot l}$ means it is the coefficient of $y$ against $a$ with another covariate $l$ in the model. The role of $\cdot$ is similar to $|$ in the conditional probabilities. In general, a linear model for $Y$ against a set of covariates $B$ is

$$E(Y \mid B) = \beta_0 + \sum_{b\in B} \beta_{by\cdot B'} x_b \tag{9}$$

where $B' := B \setminus \{b\}$. The collection of all coefficients $\beta_{By} := (\beta_{by\cdot B'})_{b\in B}$ is a vector. For simplicity, we can make $E(X_b) = 0$ for all $b \in B$ by centralising $X_b$, then $\beta_0$ is not required.

Suppose $C$ is a valid adjustment set for $A \to Y$ (and there may be continuous covariates in $A$), using model (9) with $B = C \cup \{A\}$,

$$\begin{aligned} E(Y \mid \mathrm{do}(A = a)) &= \int_{x_C \in \mathscr{X}_C} p(x_C) E(Y \mid a, x_C)\, dx_C \\ &= \int_{x_C \in \mathscr{X}_C} p(x_C) \left( a\beta_{ay\cdot C} + \sum_{c\in C} x_c \beta_{cy\cdot aC'} \right) dx_C \qquad \text{where } C' := C \setminus \{c\} \\ &= a\beta_{ay\cdot C} + \sum_{c\in C} E(X_c)\beta_{cy\cdot aC'} \\ &= a\beta_{ay\cdot C} \quad \text{because we centralised } X_c \text{ s.t. } E(X_c) = 0 \end{aligned}$$

Intervention on $A$ will affect outcome $Y$ through the regression coefficient $\beta_{ay\cdot C}$. One coefficient for one causal effect (of $A$ on $Y$) is quite sensible. The other coefficients are modelling the correlations between $X_c$ and $Y$. From the result in linear models, the least square estimator is

$$\hat{\beta}_{By} = (X_B^T X_B)^{-1} X_B^T Y$$

where $B = C \cup \{a\}$. And it has asymptotic distribution

$$N_q(0, \Sigma_B^{-1} \sigma_{yy\cdot B})$$

where $q = |C|+1$, $\Sigma_B := \mathrm{Cov}(X_B)$ is the variance-covariance matrix of $X_B$ and $\sigma_{yy\cdot B}) := \mathrm{Var}(Y|X_B)$ is the variance of $y$ when using $X_B$ as covariates. which means it is consistent. In particular, the individual coefficient

$\hat{\beta}_{ay\cdot C} \to^d N(0, \sigma_{yy\cdot aC}/\sigma_{aa\cdot C})$ (convergence in distribution)

Model (9) is a special case of *generalised linear model* with identity link function. The generalised linear model is formulated below

$$\eta(E(Y \mid B)) = \beta_0 + \sum_{b \in B} \beta_{by\cdot B'} x_c$$

where $\eta$ is the link function. For example, if $Y$ is a binary variable encoding death, then $E(Y \mid B)$ encodes the risk of death, which should be in $[0, 1]$. The link $\eta$ can be chosen to be the logit function $\eta(x) = \log(x/(1-x))$ such that

$$E(Y \mid B) = \frac{1}{1 + \exp\left(\beta_0 + \sum_{b \in B} \beta_{by\cdot B'} x_c\right)}$$

is guaranteed to fall between 0 and 1.

### 5.6.2 General Models

The linear model imposes a strict restriction (the only shape allowed is a straight line), and if $A$ has many levels, there is a large number of conditional expectations (when $A$ is continuous, there are infinitely many!), all represented using only a few parameters. So linear models are *parsimonious* in these cases. You may add more parameters to ease the restriction, for example, a polynomial model with a certain order (quadratic, cubic etc.). The predicted distribution becomes more flexible and the *bias* (expected difference between observed data and predicted value) is more controlled. However, it could raise the variance of estimation. The appropriate number of parameters ought to obtain a good *vairance-bias trade-off.*

When $A, Y$ are both continuous, all parametric models on $E(Y|A)$ are *semi-parametric.* Because they do not restrict the distribution of $Y|A$ nor the marginal distribution of $A$. Our parameters only encode part of the joint distribution of $Y, A$ through a conditional mean $E(Y|A)$.

With the (generalised) linear model, the estimation of $E(Y \mid A = a)$ at each point $a$ borrows information from all other points. However, on some occasions, values of $a$ that are far away should be irrelevant. The *kernel regression model* is required which only borrows information from points near $a$ (within an assigned *bandwidth* $h$). Often related to kernel regression is the *general additive model* [5]

$$E(Y \mid A = a) = \sum_{c \in C} f_c(X_c)$$

where $f_c$ are functions to be estimated using kernel regression and $X_c$ are the covariates.

## 5.7 Gaussian Structural Equation Model

In section 5.2, the model underlying a causal DAG $\mathcal{G}$ (with $p$ nodes, ordered in a way such that all parents of node $i$ are before it, this is the *topological ordering*), FRCISTG(NPSEM) was introduced. It assigns functions $f_i(\text{pa}_{\mathcal{G}}(i), \epsilon_i)$ to counterfactual $X_i^{x_1,\cdots,x_{i-1}} = X_i^{\text{pa}_{\mathcal{G}}(i)}$. Now assume $f_m$ are linear functions and the $\epsilon_m$ follows Gaussian distribution $N(0, d_{ii})$ for some value $d_{ii}$, i.e. the recursive generation (denoting counterfactuals by $X_i$ for simplicity) is

- (1) Initial step

$$X_1 = \epsilon_1$$

- (2) recursive step, for $i = 2, \cdots, p$,

$$X_i = \sum_{j \in \text{pa}_{\mathcal{G}}(i)} b_{ij} X_j + \epsilon_i$$

the coefficient $b_{ij}$ is non-zero iff $j \in \text{pa}_{\mathcal{G}}(i)$. This is called the *(Gaussian) Structural Equation Model*

In matrix form,

$$X_V = BX_V + \boldsymbol{\epsilon}$$

where $X_V = (X_1, \cdots, X_p)$, $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_p)$, and the matrix $B$ is

$$B := \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ b_{21} & 0 & \cdots & 0 & 0 \\ b_{31} & b_{32} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ b_{p1} & b_{p2} & \ddots & b_{p,(p-1)} & 0 \end{pmatrix}$$

. The distribution of $X_V$ is $N_p(0, \Sigma)$ where

$$\Sigma = \mathrm{Var}(X_V) = (I - B)^{-1} D (I - B)^{-T}$$

where $D = \mathrm{diag}(d_{ii})_{\in\{1,\cdots,p\}}$ is the diagonal matrix made of individual variances $d_{ii}$. (proof left as an exercise) In this section, we aim to find an easy way to evaluate the variance-covariance matrix $\Sigma$.

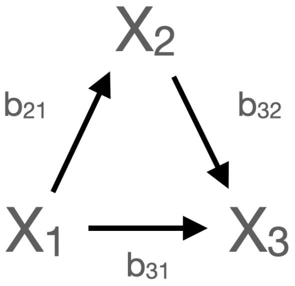


**Figure 62: A causal DAG and its SEM parameters represented on the edges**

**Example 27.** Figure 62 shows a causal DAG with three nodes and three arrows. Assume the SEM is

$$X_1 = \epsilon_1, \qquad X_2 = b_{21} X_1 + \epsilon_2, \qquad X_3 = b_{31} X_1 + b_{32} X_2 + \epsilon_3$$

where all $b_{ij}$ appeared are non-zero. In matrix form,

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ b_{21} & 0 & 0 \\ b_{31} & b_{32} & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

After rearranging, we have

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ -b_{21} & 1 & 0 \\ -b_{31} & -b_{32} & 1 \end{pmatrix}}_{=I-B} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

and inversion gives

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ b_{21} & 1 & 0 \\ b_{31} + b_{32} b_{21} & b_{32} & 1 \end{pmatrix}}_{=(I-B)^{-1}} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

we have expressed the values of $X_i$ as a linear combination of the independent errors. Note $X_3 = (b_{31} + b_{32} b_{21}) \epsilon_1 + b_{32} \epsilon_2$, and the coefficient $(b_{31} + b_{32} b_{21})$ encodes that correlation from $X_1$ to $X_3$ flows through the directed(causal) path $X_1 \to X_3$ as well as $X_1 \to X_2 \to X_3$.

There seems to be a good relation between the matrix $(I-B)^{-1}$ in the last line and $B$, indeed,

$$I+B+B^2 = \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} + \begin{pmatrix}0&0&0\\b_{21}&0&0\\b_{31}&b_{32}&0\end{pmatrix} + \begin{pmatrix}0&0&0\\0&0&0\\b_{32}b_{21}&0&0\end{pmatrix} = \begin{pmatrix}1&0&0\\b_{21}&1&0\\b_{31}+b_{32}b_{21}&b_{32}&1\end{pmatrix}$$

This is not a coincidence!

A good property of lower triangular (with zero main diagonal entries, $B_{ii}=0$) matrix is that it is *nilpotent*. i.e. $B^p = 0$. Because each multiplication by $B$ eliminates a (minor)diagonal. (see figure 63) This can be proved rigorously by writing out the matrix multiplication formula. (proof left as an exercise) From the point of linear algebra, this is because $B$ represents a linear transformation that lowers the dimension by 1 by eliminating the first base vector.

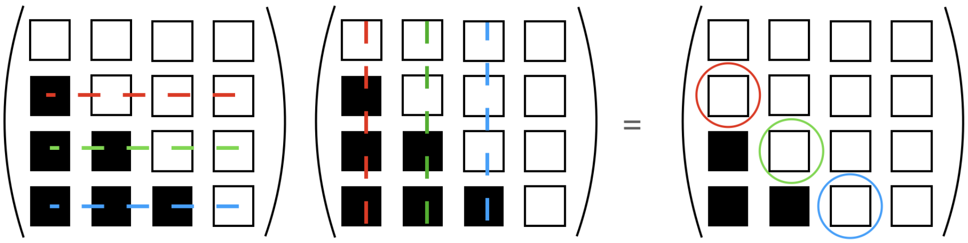


**Figure 63: Multiplication by lower diagonal matrix eliminates a diagonal**

So after $p-1$ multiplications on $B$, 0 matrix is obtained. Using the Taylor's expansion (assume it works on matrices),

$$(I-B)^{-1} = \sum_{i=0}^{\infty} B^i = I + B + B^2 + \cdots B^{p-1}$$

which provides a handy way to evaluate $\Sigma$. Further, note the $i,j$th entry of $B^n$ is

$$[B^n]_{i,j} = \sum_{k_1} \cdots \sum_{k_{n-1}} b_{ik_1} b_{k_1 k_2} \cdots b_{k_{n-1}j}$$

so it is non-zero iff $b_{ik_1}, b_{k_1k_2}, \cdots b_{k_{n-1}j}$ are all non-zero. By definition of $b_{ij}$, this means there is a directed path $j \to k_{n-1} \to \cdots \to i$. All paths of length exactly $n$ (there are $n$ edges on the path) are encoded in the non-zero entries of matrix $B^n$. Therefore, the $i,j$ entry of $(I-B)^{-1}$ encodes $b$ coefficients along all paths from $X_i$ to $X_j$. And if $[(I-B)^{-1}]_{i,j} = 0$, there is no directed path between $i$ and $j$.

**Example 28.** Study Figure 62 again. For simplicity assume $D = I$(unit, equal variance) We know the correlation between $X_1$, $X_3$ only flows through the two causal(directed) paths mentioned above. So the covariance $\mathrm{Cov}(X_1,\, X_3) = b_{31} + b_{32}b_{21}$. Indeed, if you calculate $\Sigma = (I-B)^{-1}(I-B)^{-T}$, the $1,3$ entry is $b_{31} + b_{32}b_{21}$.

What about the covariance between $X_2$ and $X_3$? There is a direct causal path $X_2 \to X_3$, a confounding effect from $X_2 \leftarrow X_1 \to X_3$ and the covariance linking back to the parent of $X_2$, $X_1$, then through the causal path to $X_3$, which is $X_2 \leftarrow X_1 \to X_2 \to X_3$. Correlation is symmetric, so considering paths from $X_2$ to $X_3$ is enough. Indeed, reading from the matrix $\Sigma = (I-B)^{-1}(I-B)^{-T}$,

$$\mathrm{Cor}(X_1,\, X_3) = b_{32} + b_{21}b_{31} + b_{21}^2 b_{32}$$

the last two terms are all introduced by the confounder $X_1$.

In general, we define paths in the last paragraph as *treks* [27].

**Definition 34** (Trek)**.** A trek from node $i$ to $j$ is either $i \to j$ (when $i = j$, we denote this path simply as $i$), or an ORDERED pair of directed paths $(\pi_l, \pi_r)$ where $\pi_l$ is a path from $k$ to $i$ and $\pi_r$ is a path from $k$ to $j$. The node $k$ (a confounder of $i$, $j$) is called the *source*, and $\pi_l, \pi_r$ are called left and right *side* of the trek respectively. In the second case $((\pi_l, \pi_r))$, the trek is called a *two-sided trek with source k*.

This definition formalises the concept of ‘a path(may not be directed) where correlation can flow’, including both the causal path and other back-door paths not blocked by a collider.

**Definition 35** (trek covariance)**.** The trek covariance of $\tau = (\pi_l, \pi_r)$ is

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

i.e. product of all $b$ values along the left and right sides of source $k$ multiplied by the variance of error $\epsilon_k$, $d_{kk}$.

**Example 29.** Study Figure 62 for a third time. To find the variance of $X_3$ (covariance between $X_3$ and itself), we need to find all treks from $X_3$ to $X_3$, which are as follows (I have underlined the source)

- no source: $X_3$
- source $X_1$: $X_3 \leftarrow \underline{X_1} \rightarrow X_3$, $X_3 \leftarrow \underline{X_1} \rightarrow X_2 \rightarrow X_3$, $X_3 \leftarrow X_2 \leftarrow \underline{X_1} \rightarrow X_3$, and $X_3 \leftarrow X_2 \leftarrow \underline{X_1} \rightarrow X_2 \rightarrow X_3$
- source $X_2$: $X_3 \leftarrow \underline{X_2} \rightarrow X_3$

when finding treks from a node to itself, there are many repetitions of nodes. After all, $\text{Var}(X_3) = \text{Cov}(X_3, X_3)$ so there is supposed to be a lot of square terms. And summing the trek covariance yields

$$\text{Var}(X_3) = 1 + b_{32}^2 + b_{31}^2 + 2b_{31}b_{21}b_{32} + b_{21}^2 b_{32}^2$$

To formalise what we have done in the last two examples, the following theorem is presented.

**Theorem 5.11** (Trek Rule)**.** *For causal DAG $\mathcal{G}$ with Gaussian SEM, the covariance between $X_i, X_j$ is given by the sum of treks covariance, i.e.*

$$\sigma_{ij} := Cov(X_i, X_j) = \sum_{\tau\in\mathcal{T}_{ij}} c(\tau)$$

*where $\mathcal{T}_{ij}$ is the set of all treks from $i$ to $j$*

*Proof.* Proof by induction on the number of variables $p := |\mathcal{G}|$. Use the topological ordering of nodes (so that the last node $X_p$ has no child) When $p = 1$, there is only one trek $X_1$, and indeed $\sigma_{11} = d_{11}$.

Assume the trek rule is true for any graph with less than $p$ nodes. i.e. it is true on $\mathcal{G}_{V\setminus\{p\}}$. We have to show trek rule is true for $\text{Cov}(X_i,\, X_p)$ where $i \neq p$, and $\text{Cov}(X_p,\, X_p)$. By SEM, $X_p = \sum_{j\in\text{pa}(p)} b_{pj}X_j + \epsilon_p$. By assumption of SEM, all nodes in $\text{pa}(p)$ are only dependent through $\epsilon_1, \cdots, \epsilon_{p-1}$, which means $\epsilon_p$ is independent of all nodes $X_j$ with $j \in \text{pa}(p)$. So

$$\text{Cov}(X_i,\, X_p) = \sum_{j\in\text{pa}(p)} b_{pj}\text{Cov}(X_i,\, X_j)$$

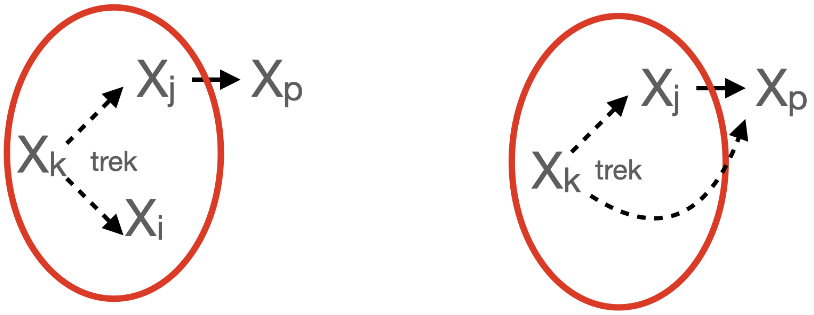


**Figure 64: Extension of treks on the reduced graph(red circle). Note it has $p-1$ nodes so covariance between any two variables is the sum of trek covariances by inductive assumption**

As illustrated in the left of Figure 64, any trek from $X_i$ to $X_j$(a parent of $X_p$) naturally extends to a trek from $X_i$, $X_p$ by the additional edge $X_j \to X_p$. So

$$\begin{aligned}
\mathrm{Cov}(X_i,\, X_p) &= \sum_{j\in \mathrm{pa}(p)} b_{pj} \sum_{\tau\in\mathcal{T}_{ij}} c(\tau) \quad \text{by inductive assumption} \\
&= \sum_{\tau\in\mathcal{T}_{ij}} \sum_{j\in \mathrm{pa}(p)} b_{pj} c(\tau) \\
&= \sum_{\tau'\in\mathcal{T}_{ip}} c(\tau') \quad \text{where } \tau' = \tau \cup \{j \to p\}
\end{aligned}$$

To the right of figure 64, any trek from $X_p$ to $X_p$ other than $X_p$ itself involves one of its parent $X_j$.

$$\begin{aligned}
\mathrm{Cov}(X_p,\, X_p) &= \sum_{j\in \mathrm{pa}(p)} b_{pj}\mathrm{Cov}(X_p, X_j) + \mathrm{Cov}(X_p, \epsilon_p) \\
&= \sum_{j\in \mathrm{pa}(p)} b_{pj} \sum_{\tau\in\mathcal{T}_{pj}} c(\tau) + \mathrm{Cov}(\epsilon_p, \epsilon_p) \qquad \text{by the last step} \\
&= \sum_{\tau\in\mathcal{T}_{pj}} \sum_{j\in \mathrm{pa}(p)} b_{pj} c(\tau) + d_{kk} \\
&= \sum_{\tau'\in\mathcal{T}_{pp}\backslash\{p\}} c(\tau') + c(p) \quad \text{where } c(p) \text{ is the covariance of trivial trek } p \\
&= \sum_{\tau'\in\mathcal{T}_{pp}} c(\tau')
\end{aligned}$$

□

## 5.8 Efficient Adjustment

Finally, we come back to the study of adjustment sets. In section 5.5, we discussed some criteria for valid adjustment sets. Valid adjustment sets are not unique, and in reality, we need the one with the best performance and the lowest computational cost.

If there is no random error (the data of the whole population is available), a smaller adjustment set is always better as it is easier to calculate. However, in reality, there are random errors. Minimal adjustment sets(minimal cardinality) need to be constructed whilst maintaining a decent efficiency[11][23].

Recall that under the Gaussian(assuming covariates $T, \boldsymbol{X}_C$ are Gaussian distributed) linear model, the conditional expectation is

$$E(Y|T, \boldsymbol{X}_C) = \beta_{ty\cdot C}T + (\boldsymbol{\beta}_{Cy\cdot t})^T\boldsymbol{X}_C$$

where absence of intercept means the data is centralised ($E(\boldsymbol{X}_C) = 0$, $E(T) = 0$). The causal effect of $T$ on $Y$ under stratification by $C$(conditioning on $C$) is $\beta_{ty\cdot C}$.
Considering the covariance between treatment $T$ and outcome $Y$,

$$\mathrm{Cov}(T,\, Y \,|\, \boldsymbol{X}_C) = \mathrm{Cov}(T,\, \beta_{ty\cdot C}T + (\boldsymbol{\beta}_{Cy\cdot t})^T\boldsymbol{X}_C \,|\, \boldsymbol{X}_C) = \beta_{ty\cdot C}\mathrm{Cov}(T, T \,|\, \boldsymbol{X}_C)$$

so $\beta_{ty\cdot C} = \sigma_{ty\cdot C}/\sigma_{tt\cdot C}$. This is the continuous version of "causal effect equals correlation under stratification by $C$":

$$p(y^t \,|\, \boldsymbol{x}_C) = p(y \,|\, \mathrm{do}(T = t), \boldsymbol{x}_C) = p(y \,|\, t, \boldsymbol{x}_C)$$

(i.e. the conditional exchangeability $Y^t \perp T \,|\, \boldsymbol{X}_C$ holds)

The least-square estimator of causal effect $\beta_{ay\cdot C}$ has asymptotic distribution

$$\hat{\beta}_{ty\cdot C} \to^d N(0, \sigma_{yy\cdot tC}/\sigma_{tt\cdot C})$$

so we can compare two adjustment sets $C, D$ by comparing their asymptotic variances. We prefer $C$ over $D$ if

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

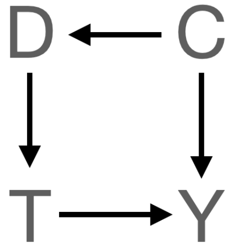


**Figure 65: Comparison of two adjustment methods $C$ and $D$**

**Example 30.** For the causal DAG in Figure 65, both conditioning on $C$ and $D$ blocks the back-door path $T \leftarrow D \leftarrow C \rightarrow Y$. So $\{C\}$ and $\{D\}$ are both valid adjustment sets. By intuition, adjusting $C$ is better as it is the source of confounding (cause of both $T$ and $Y$) in this case. $D$ is just a mediator of $C$'s effect on treatment $T$ and hold limited information on the outcome $Y$, in terms of independence, $D \perp Y \,|\, T, C$. i.e. if the treatment is fixed (intervened), all information on $Y$ is contained in $C$. So conditioning on $C$ should give a smaller conditional variance of $Y$ compared to continuing on $D$. i.e.

$$\sigma_{yy \cdot tC} = \sigma_{yy \cdot tCD} \leq \sigma_{yy \cdot tD}$$

Similarly, $C \perp T \,|\, D$, so $D$ holds more information on $T$ than $C$, which results in

$$\sigma_{tt \cdot D} = \sigma_{tt \cdot CD} \leq \sigma_{tt \cdot C}$$

so asymptotic variance of $\hat{\beta}_{ty \cdot C}$ will be smaller than that of $\hat{\beta}_{ty \cdot D}$. An interpretation of this example is that the study of the asymptotic variance of regression coefficients is an outcome-oriented approach in the sense that smaller asymptotic variance yields a more stable prediction of the conditional expectation of $Y$ (which equals the conditional causal effect if the condition is a valid adjustment set). So adjustment sets carrying more information about the outcome $Y$ are preferable than those carrying information about treatment $T$.

To formalise "more information = lower variance", the following lemma[8] is used

**Lemma 5.12** (Deletion of covariates)**.** *Suppose $Y, X, S$ are jointly Gaussian distributed, where $S$ could be a set. The following equation is true*

$$\sigma_{yy \cdot xs} = \sigma_{yy \cdot x} - \boldsymbol{\beta}_{ys \cdot x} \Sigma_{ss \cdot x} (\boldsymbol{\beta}_{ys \cdot x})^T$$

Since the inner product $\boldsymbol{\beta}_{ys \cdot x} \Sigma_{ss \cdot x} (\boldsymbol{\beta}_{ys \cdot x})^T \geq 0$, we have $\sigma_{yy \cdot xs} \leq \sigma_{yy \cdot x}$. i.e. conditioning on more covariates yields lower variance.

The example above can be generalised to a criterion of comparing two adjustment sets, and we assume $C, D$ may not be disjoint.

**Proposition 5.13.** *Given adjustment sets $C$, $D \subseteq V$ ($C, D$ do not contain $\{T, Y\}$) Define $C' = C \setminus D$(unique elements of $C$) and $D' = D \setminus C$(unique elements of $D$), if*

*$Y \perp X_{D'} \,|\, X_C, T$ (C has more information on outcome) $Y \perp X_{C'} \,|\, X_D$ (D has more information on treatment)*

*then the estimated causal effect has a smaller asymptotic variance when adjusted by $C$ than adjusted by $D$, i.e.*

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

*Proof.* Since $C \cup D' = C' \cup D$, independence $Y \perp X_{D'} \,|\, X_C$, $T$ implies

$$\sigma_{yy \cdot tC} = \sigma_{yy \cdot tCD'} = \sigma_{yy \cdot tC'D} \leq \sigma_{yy \cdot tD}$$

where the inequality is by lemma 5.12.
Similarly, independence $Y \perp X_{C'} \,|\, X_D$ implies

$$\sigma_{tt \cdot D} = \sigma tt \cdot C'D = \sigma_{tt \cdot CD'} \leq \sigma_{tt \cdot C}$$

□

Proposition 5.13 gives us a rule for adding or removing nodes from a valid adjustment set $C$[6].

**Corollary 3.** *Given valid adjustment set $C$ and $P \in pa_{\mathcal{G}}(T) \cap C$, $R \in pa_{\mathcal{G}}(Y)$, then*

- *$C \setminus \{P\}$ is preferable than $C$ (in terms of asymptotic variance) if $C \setminus \{P\}$ is a valid adjustment set*
- *$C \cup \{R\}$ is preferable than $C$ (in terms of asymptotic variance) if $C \cup \{R\}$ is a valid adjustment set*

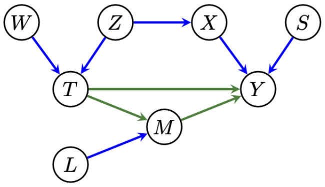


**Figure 66: A causal diagram with two causal paths**

**Example 31.** Consider the causal diagram in Figure 66 where the effect of treatment $T$ on $Y$ is through direct causation and also a mediator $M$. (there are two causal paths) So the causal nodes are $\text{cn}(\mathcal{G}) = \{M, Y\}$. There is only one back-door path $T \leftarrow Z \rightarrow X \rightarrow Y$. So $\{Z\}$ and $\{X\}$ are valid adjustment sets. To estimate the conditional expectation of $Y$ better, $\{X\}$, which is closer to $Y$ is preferable. (this is also the result of 5.13)
By corollary 3, adding other parents of $Y$ would also lower the asymptotic variance, so expand the set to $\{X, S, L\}$ (of course, things in the forbidden set cannot be added, which are $T$ and $M$ in this example). Note, $L$ benefits the estimation of mediator $M$, which is a parent of $Y$. So $L$ is also added. We should have arrived at an optimal set in the sense that the asymptotic variance is lowest across all valid adjustment set.

The construction of the optimal set above aims to gain more information on the outcome $Y$ and throw unnecessary information on treatment $T$ while preserving validity. This is generalised to an *O-set*[6].

**Definition 36** (O-set)**.** Given nodes $T, Y \in V$, the O-set is

$$O_{\mathcal{G}}(T \to Y) = \text{pa}_{\mathcal{G}}(\text{cn}_{\mathcal{G}}(T \to Y)) \setminus \text{forb}_{\mathcal{G}}(T \to Y) = \text{pa}_{\mathcal{G}}(\text{cn}_{\mathcal{G}}(T \to Y)) \setminus (\text{cn}_{\mathcal{G}}(T \to Y) \cup \{T\})$$

*Remark.* Although we used $T \to Y$ in the definition, the O-set is well-defined as long as $Y$ is a descendant of $T$. The second inequality is true because strict descendants of causal nodes are clearly not in $\text{pa}_{\mathcal{G}}(\text{cn}_{\mathcal{G}}(T \to Y))$.

O-set is indeed the (asymptotically) optimal set [6]

**Theorem 5.14.** *If $\mathcal{G}$ is causal DAG with variables $T$, $Y$ and $Y \in de_{\mathcal{G}}(T)$. Assume there exists a valid adjustment set for $T, Y$. Then the O-set is a valid adjustment set, and the asymptotic variance of $\hat{\beta}_{ty \cdot O}$ is minimal over all other valid adjustment sets for $T, Y$*

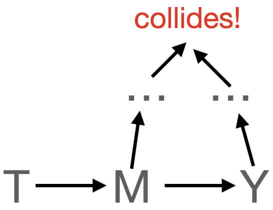


**Figure 67: A back-door path without parent of causal nodes ($M, Y$ in this case) must have a collider**

*Proof.* The O-set is valid because any open back-door path for $T, Y$ must involve at least one parent of causal nodes (otherwise, there must be a collider on the path, see Figure 67), and it does not contain any forbidden node by definition. So O-set satisfies the generalised adjustment criterion.

To prove optimality, for any valid adjustment set $Z$. Define $O' = O \setminus Z$, $Z' = Z \setminus O$. Aim to prove $Y \perp_d Z' \,|\, O \cup \{T\}$ and $O \perp T \,|\, Z'$. Then proposition 5.13 yields that asymptotic variance of $\hat{\beta}_{ty \cdot O}$ is smaller than that of $\hat{\beta}_{ty \cdot Z}$.

**Target 1**: $Y \perp_d Z' \,|\, O \cup \{T\}$:
Note $Z \cap \text{forb}_{\mathcal{G}}(T \to Y) = \emptyset$ as $Z$ is valid. By definition of $O$, it includes all parents of the causal nodes (except $T$ and causal nodes themselves). So the only nodes left in $Z'$ are either parents of $T$ or descendants of $T$ (that are not descendants of any causal node).

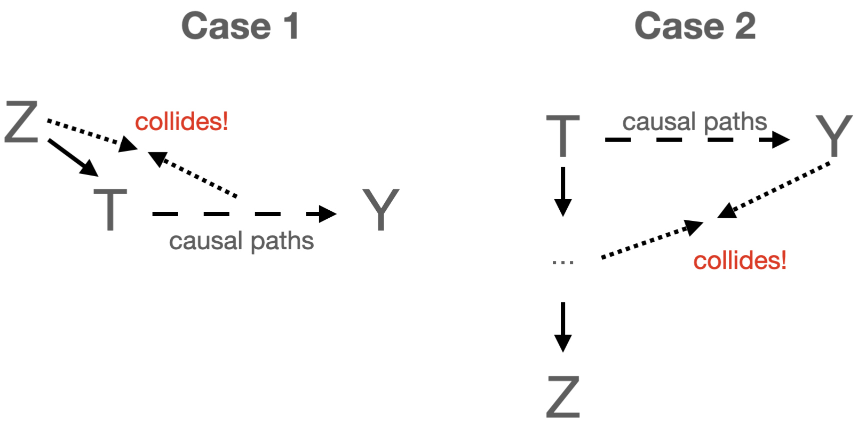


**Figure 68: Possible paths from $Z \in Z'$ to $Y$**

In the first case(left of Figure 68), $Z \in Z'$ is a parent of $T$, but it cannot be a parent of anything else on the causal paths (as those are in the O-set). So any path to $Y$ is either through $T$ or through a descendant of a causal node. We have illustrated in Figure 67 that the second scenario must have a collider. So the path is not open.
In the second case (right of Figure 68), a path from $Z$ (descendant of $T$) to $Y$ must go through $T$ or a node between $Z$ and $T$. In the second scenario, to ensure $Z$ is not a descendant of $Y$, there must be a collider on this path.

**Target 2**: $O \perp T \,|\, Z'$:
Suppose for contradiction there is an open path $\pi_O$ from $O \in O'$ to $T$ not going through $Z$. Since $O$ is not a parent of $T$ by definition of O-set, $\pi_O$ cannot be a directed path from $O$ to $T$. Meanwhile, $O$ must be a parent of a causal node, so we can concatenate $\pi_O$ with any directed path $\pi_C$ from $O$ to $Y$. The path $(\pi_O, \pi_C)$ is a back-door path as $O$ should not be on any causal path (forbidden).

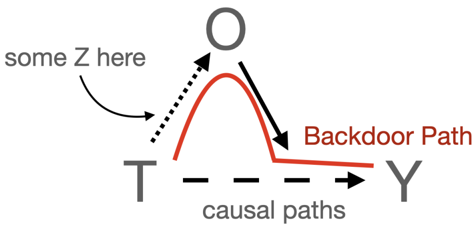


**Figure 69: Possible path from $O \in O'$ to $T$**

But $Z$ is a valid adjustment set, so it must block this back-door path on $\pi_O$ (not on $\pi_C$ as it contains the causal nodes $\text{cn}(T \to Y)$). Contradiction. □

### 5.8.1 Forbidden Projection

Optimal adjustment (in terms of reducing asymptotic variance) favours nodes that carry more information on the outcome $Y$, which are preferably direct parents of all causal nodes. However, forbidden nodes should be avoided. So we can construct a new graph and shrink all the forbidden nodes without disrupting other structures.

**Definition 37** (Latent Projection[25])**.** Given DAG $\mathcal{G}$ with nodes $V \cup L$ where $L$ contains all the latent variables (e.g. unmeasured variables) The latent projection of $\mathcal{G}$ over $V$ (it removes $L$) is a new graph $\tilde{\mathcal{G}}^{V,L}$ where

- $v \to_{\tilde{\mathcal{G}}} w$ iff there is a directed path in $\mathcal{G}$ from $v$ to $w$ with all nodes between in $L$. i.e. a directed edge in $\tilde{\mathcal{G}}$ means causation (with possible latent mediators)
- $v \leftrightarrow_{\tilde{\mathcal{G}}} w$ iff there is a two-sided trek from $v$ to $w$ with all nodes in-between in $L$. i.e. there is an extra correlation following between $v$, $w$ other than causation (with possible latent confounders etc.)

*Remark.* $\tilde{\mathcal{G}}^{V,L}$ is called an *acyclic directed mixed graph*(ADMG) because it has both directed and bi-directed edges.

**Definition 38** (Forbidden Projection[26])**.** Let $L = \text{forb}_{\mathcal{G}}(T \to Y) \setminus \{T, Y\}$, the forbidden projection is the latent projection of $\mathcal{G}$ over $L$ onto $V \setminus L$. i.e. $\tilde{\mathcal{G}}^{V\setminus L,L} = \tilde{\mathcal{G}}$.

*Remark.* Note we are projecting over the causal nodes. If there is a two-sided trek between $v$ to $w$ with all in-between nodes in $L$, then it must be a directed path, contradicting the definition of a two-sided trek. So no bi-directed edges are in $\tilde{\mathcal{G}}^{V\setminus L,L}$.

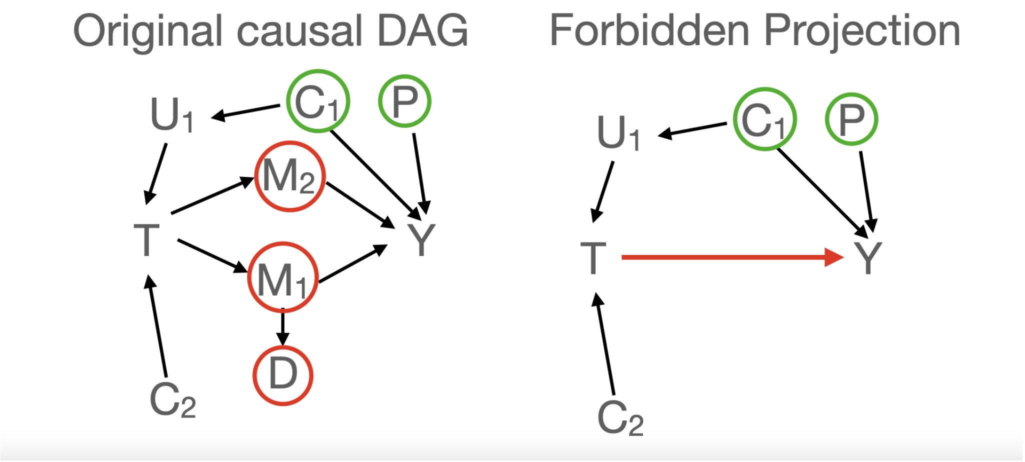


**Figure 70: The forbidden projection of a causal DAG**

**Example 32.** Figure 70 shows a causal DAG with two mediators $M_1$, $M_2$ and one confounder $C_1$. All the forbidden nodes (except $T, Y$) are marked in red circles. As you can see, all red circles are projected out in the forbidden projection. I have deliberately marked the edge $T \to Y$ in red to represent there are hidden forbidden nodes between them in the original graph. The optimal set is marked in green circles. After projection, green circles are exactly the parents of $Y$.

**Theorem 5.15.** *Suppose there exists a valid adjustment set for causal DAG $\mathcal{G}$, and $\tilde{G}$ is the forbidden projection for $T, Y$. Then*

$$O_{\mathcal{G}}(T \to Y) = pa_{\tilde{\mathcal{G}}}(Y) \setminus \{T\}$$

[26] also proposed efficient graphical and non-graphical algorithms for constructing the O-set.

## 6 Ending

The auxiliary notes on Graphical Models finish here. Thank you very much for reading.
Please email daniel.kansaki@outlook.com or yuhang.lin@new.ox.ac.uk if you find any typo or have suggestions for improvements.

# References

[1] Columbia University Irving Medical Centre. Difference-in-difference estimation, Mar 2023.

[2] E. Cui, E. Weng, E. Yan, and J. Xia. Robust leaf trait relationships across species under global environmental changes. *Nat. Commun.*, 11(1):2999, 2020.

[3] Adnan Darwiche. *Modeling and reasoning with Bayesian Networks*. Cambridge University Press, 2009.

[4] David Edwards. *Introduction to graphical modelling*. Springer, 2000.

[5] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310, 1986.

[6] Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, March 2019.

[7] Miguel A. Hernan and James M. Robins. *Causal inference: what if*. Taylor and Francis, 2020.

[8] Manabu Kuroki and Zhihong Cai. Selection of identifiability criteria for total effects by using path diagrams, 2012.

[9] Steffen L. Lauritzen. *Graphical models*. Clarendon Press, 1996.

[10] Steffen L. Lauritzen, A. Philip Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed markov fields. *Networks*, 20:491–505, 1990.

[11] XAVIER DE LUNA, INGEBORG WAERNBAUM, and THOMAS S. RICHARDSON. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875, 2011.

[12] Judea Pearl. Chapter 3 - markov and bayesian networks: Two graphical representations of probabilistic knowledge. In Judea Pearl, editor, *Probabilistic Reasoning in Intelligent Systems*, pages 77–141. Morgan Kaufmann, San Francisco (CA), 1988.

[13] JUDEA PEARL. Causal diagrams for empirical research. *Biometrika*, 82(4):702–710, 1995.

[14] Judea Pearl. *Causality models, reasoning, and inference*. Cambridge University Press, 2000.

[15] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018.

[16] Judea Pearl and Azaria Paz. *GRAPHOIDS: Graph-Based Logic for Reasoning about Relevance Relations OrWhen Would x Tell You More about y If You Already Know z?*, page 189–200. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022.

[17] Thomas S. Richardson and James M. Robins. Single world intervention graphs : A primer. 2013.

[18] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.

[19] Donald B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, 2006.

[20] Donald B Rubins. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, page 365–382, 2001.

[21] Ilya Shpitser, Tyler VanderWeele, and James M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*, Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010, pages 527–536. AUAI Press, 2010. Copyright: Copyright 2020 Elsevier B.V., All rights reserved.

[22] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

[23] Johannes Textor and Maciej Liskiewicz. Adjustment criteria in causal diagrams: An algorithmic perspective. *CoRR*, abs/1202.3764, 2012.

[24] Tyler J. VanderWeele and Onyebuchi A. Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, 22(1):42–52, 2011.

[25] TS Verma and Judea Pearl. Equivalence and synthesis of causal models. *in Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, page 220–227, 1990.

[26] Janine Witte, Leonard Henckel, Marloes H. Maathuis, and Vanessa Didelez. On efficient adjustment in causal graphs, 2020.

[27] Sewall Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.

[28] Eric P. Xing. Lecture 2: Directed gms: Bayesian networks - cmu school of computer science, 2015.

[29] Eric P Xing. Lecture 4: Exact inference - cmu school of computer science, 2017.