Decision-Making & Modelling Under Uncertainty (DMU)

Oliver Maclaren (oliver.maclaren@auckland.ac.nz)

[10 lectures/tutorials]

- Decision-making under uncertainty [5/10]
  - Basic concepts
  - Risk, probability, utility
  - Statistical: extended setup
    - formulation & empirical risk approx.
    - minimax & Bayes
  - Tutorial sheet
- Modelling under uncertainty { models of risk & intervention [5/10]
  - probability, graphical models, & independence
  - causal interpretations of graphical models
  - stochastic process models (esp. Markov)
  - simulation & estimation tools
  - Tutorial sheet

Lecture 5: Modelling under risk

→ Representation & independence for probabilistic models

In the next few lectures we will look at models of, or incorporating, risk/probability

— these may represent 'frequentist' probabilities ('out there in the world') [or] 'belief' (Bayesian) probabilities

— Using Bayes' theorem ≠ being a 'Bayesian'

— 'Bayesian' means using probability to represent belief

→ See Wasserman →

From Larry Wasserman (author of 'All of Statistics'):

1. Some Obvious (and Not So Obvious) Statements

Before I go into detail, I'll begin by making a series of statements.

Frequentist Inference is Great For Doing Frequentist Inference.
Bayesian Inference is Great For Doing Bayesian Inference.

Frequentist inference and Bayesian Inference are defined by their goals, not their methods.

A Frequentist analysis need not have good Bayesian properties.
A Bayesian analysis need not have good frequentist properties.

Bayesian Inference $\neq$ Using Bayes Theorem

Bayes Theorem $\neq$ Bayes Rule

Bayes Nets $\neq$ Bayesian Inference

Frequentist Inference is not superior to Bayesian Inference.
Bayesian Inference is not superior to Frequentist Inference.
Hammers are not superior to Screwdrivers.

Confidence Intervals Do Not Represent Degrees of Belief.
Posterior Intervals Do Not (In General) Have Frequency Coverage Properties.

Saying That Confidence Intervals Do Not Represent Degrees of Belief Is Not a Criticism of Frequentist Inference.
Saying That Posterior Intervals Do Not Have Frequency Coverage Properties Is Not a Criticism of Bayesian Inference.

Some Scientists Misinterpret Confidence Intervals as Degrees of Belief.
They Also Misinterpret Bayesian Intervals as Confidence Intervals.

Mindless Frequentist Statistical Analysis is Harmful to Science.
Mindless Bayesian Statistical Analysis is Harmful to Science.

See:

https://normaldeviate.wordpress.com/2012/11/17/what-is-bayesianfrequentist-inference/

→ The tools we discuss may be useful for either (or something else!)

Probability models: representation & computation

[See summary sheet for background]

→ In principle a risk/probability model can be captured by a joint distribution over a collection of random variables

(see later for stochastic process models)

→ However, in practice it is difficult to represent & compute directly with joint distributions

Aside: densities & mass functions

- For discrete random variables, each outcome has finite probability & so eg $\mathbb{P}(X=1)$ makes sense
- The function giving the probabilities of each outcome, $\mathbb{P}(X=x)$, is the probability mass function (pmf) & we write $\boxed{p(x) = \mathbb{P}(X=x)}$
- For continuous random variables each outcome typically only has 'infinitesimal' probability & for finite probabilities we have to consider eg intervals: $\mathbb{P}(1 \leq X \leq 2)$ etc.

→ This is given by $\int_1^2 p(x)\,dx$ etc. for probability density function (pdf) $p(x)$ (or $p_X(x)$ or $f_X(x)$)

Aside: densities & mass functions

We can relate these considering (semi-formally) the probability mass of a continuous outcome as

'$\mathbb{P}(X=x) = p(x)\,dx$' $\left(\approx \int_x^{x+dx} p(z)\,dz\right)$

pdf, differential

The discrete case is of course just:

'$\mathbb{P}(X=x) = p(x)$'

pmf

Importantly: In our context, most key relationships can be expressed in terms of '$p(x)$' for both continuous & discrete vars, regardless of whether discrete (pmf) or continuous (pdf).

→ We will hence call both of these 'probability functions' & let context decide if $p(x)$ is a pdf or pmf.

Illustration : representation troubles

Consider N binary random variables

X_i , $i = 1, \ldots, N$

ie each X_i can take values in $\{0, 1\}$

The joint distribution over these is given by

$$P(X_1 = x_1, \ldots X_N = x_N) = P(x_1, x_2, \ldots, x_N)$$

which requires specifying 2^N possible probabilities (one per input combination)

→ hard as N grows!

Thus, even representation (writing down a model) is hard for probability models with many variables

→ motivates →

Independence for 'modularity'

- Random variables X & Y are called independent if

$$\boxed{P(x, y) = P(x)P(y)}$$

$$\Leftrightarrow$$

$$\boxed{P(y|x) = P(y)}$$

$$\Leftrightarrow$$

$$\boxed{P(x|y) = P(x)}$$

(recall $P(X = x) = P(x)dx$ notation)

We write this as $\boxed{Y \perp\!\!\!\perp X}$ which is equivalent to $\boxed{X \perp\!\!\!\perp Y}$

- This also applies to sets of random vectors like $\{X_1, X_2, \ldots X_n\}$

→ We can write $X_1 \perp\!\!\!\perp \{X_2, \ldots, X_n\}$ etc to express that X_1 is independent of the other RVs

→ Independence assumptions make models more modular & hence tractable

(see illustration... but first →

Mutual independence of $X_1, X_2, \ldots, X_n$

is expressed as:

$$X_i \perp\!\!\!\perp \{X_j\}_{j \neq i}, \text{ for all } x_i$$

where $\{X_j\}_{j \neq i}$ is the set of all RVs except X_i.

Random variables X & Y are called

'conditionally independent given Z'

if $P(y|x,z) = P(y|z)$

or, equivalently:

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

or, equivalently:

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

!!

This is written: $X \perp\!\!\!\perp Y \mid Z$

$(\Leftrightarrow Y \perp\!\!\!\perp X \mid Z)$

We can also consider eg $X \perp\!\!\!\perp Y \mid \{Z_1, \ldots Z_n\}$

→ X & Y independent given the set of random vars $Z_1, \ldots Z_n$

Modularity: illustration

Suppose in our example that the N binary variables are mutually independent

then:

$$P(x_1, x_2, x_3, \ldots x_N)$$

$$= P(x_1) \cdot P(x_2) \cdot \ldots P(x_N)$$

$$= \prod_{i=1}^{N} P(x_i)$$

~ product version of sum $\sum_{i=1}^{n}$

→ only need $2N$ values to be given instead of 2^N. Much less when N big!

→ While mutual independence is rare, in practice conditional independences are not!

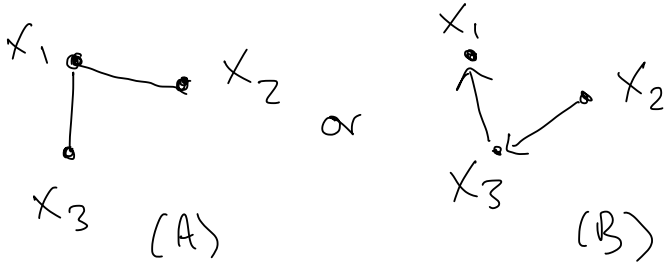
→ also leads to computational improvements

→ also conceptual/modelling advantages

Representing conditional independences

→ Graphical models are a popular way to represent conditional independencies

→ These use graphs consisting of vertices representing variables & edges representing (in)dependencies

eg 

→ Two common types of graph are
- 'undirected': edges have no direction (A)
- 'directed': edges have a direction (B)

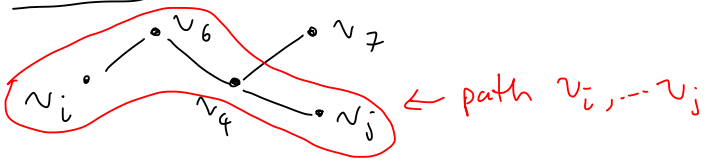
Aside:

A graph $G = (V, E)$ is defined formally as a structure containing

→ a set of vertices V &

→ a set of edges E between vertices.

- We also call vertices 'nodes'
- We can consider an edge $e \in E$ as a tuple (v_1, v_2) for some $v_1, v_2 \in V$
- A path between two vertices v_i, v_j is a sequence of vertices $v_i, \dots, v_j$, starting at v_i & ending at v_j, such that there is an edge between each consecutive pair of vertices

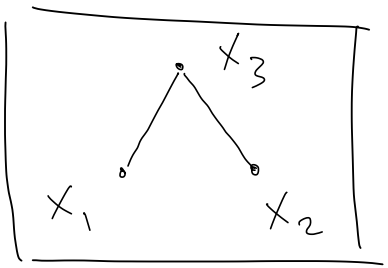


Undirected graphs : independence interpretation

These represent purely probabilistic (cf causal - see later) independencies using the rule:

Given a set of variables $X_1, \ldots X_N$, we do not draw an edge between a pair X_i & X_j if $\boxed{X_i \perp\!\!\!\perp X_j \mid \text{(the rest)}}$

→ eg



graph

represents $\boxed{X_1 \perp\!\!\!\perp X_2 \mid X_3}$ independence

ie $\boxed{P(x_1 \mid x_2, x_3) = P(x_1 \mid x_3)}$ simplification of prob. function.

etc

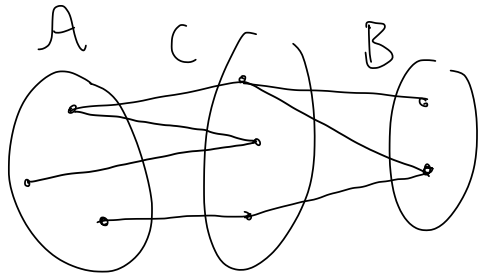
Undirected separation (u-separation)

The graph property of u-separation relates sets of vertices (nodes) in an undirected graph as follows:

Given three sets of vertices $A, B, C \subseteq V$, we say C separates A & B if all paths from elements of A to B pass through C

---

eg:



Sets of variables & independence:

The graph property of separation translates to the probabilistic property of independence:

If:

$\{Z_1, \ldots, Z_p\}$ separates $\{X_1, \ldots, X_n\}$ & $\{Y_1, \ldots Y_m\}$

then:

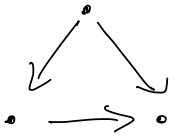
$$\{X_1, \ldots, X_n\} \perp\!\!\!\perp \{Y_1, \ldots, Y_m\} \mid \{Z_1, \ldots, Z_p\}$$

Called the 'global Markov property'

What about directed graphs?

→

Directed graphical models



- Also called 'Bayesian networks' (misnomer: not really 'Bayesian')
- Designed to incorporate information about both 'causality', &/or directional ideas, & probability

Recall: 'correlation ≠ causation'

↑ probability concept ↑ not probability concept?

- First we will look at how they encode (probabilistic) conditional independence
  - ↳ different to undirected graphs
  - ↳ can use just as alternative tool for representing conditional independencies
  - ↳ or, also as causal modelling tool

Directed graphical models

- Edges now have directions (are arrows): $\rightarrow$
- We will restrict attention to acyclic directed graphs or **DAGs** (directed acyclic graphs)

DAG? No directed cycles, where:

→ A cycle is a path from a node back to itself

→ A directed cycle is a cycle of directed edges where they all have the 'same' direction along the path

eg

directed cycle

no directed cycle

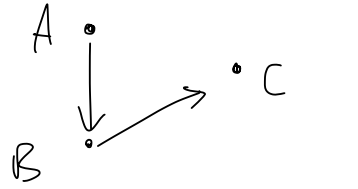
Directed independence?

The idea of a DAG is to encode a type of conditional independence representing an 'ordered Markov factorisation' of the probability model

---

First define a 'predecessor' or 'ancestor' of a vertex v as any other vertex occurring 'before' v on a directed path connecting both vertices

Eg.



A has no ancestors
B has A as anc.
C has both A & B as ancestors

A parent is an immediate ancestor (single edge path)

→ Excercise: Define descendents & children in obvious way!

Given these definitions, a DAG directly encodes conditional independences of the form:

$$X \perp \text{Pred}(X) \setminus \text{Pa}(X) \mid \text{Pa}(X)$$

where
- $\text{Pred}(X)$ = 'Predecessors of X'
- $\text{Pa}(X)$ = 'Parents of X'
- $\text{Pred}(X) \setminus \text{Pa}(X)$ = 'Non parent predecessors'

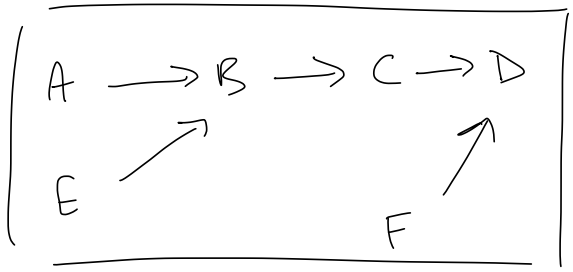
ie 'X is independent of its non-parent predecessors, given its parents'

(only the immediate predecessors, ie parents, matter)

$$\Rightarrow p(x \mid \text{pred}(x)) = p(x \mid \text{pa}(x))$$

[where $\text{pa}(x)$ stands for 'values of parents of X' etc.]

Eg:



Encodes independences such as:

$$C \perp \{A, E\} \mid B$$

ie

$$P(c \mid a, b, e) = P(c \mid b)$$

etc.

Chain rule of probability

Given a collection of N random variables and any ordering of them : $X_1, X_2, \dots X_N$

we can always write the joint distribution as

$$P(x_1, x_2, \dots, x_n) = P(x_n | x_{n-1}, \dots x_1) P(x_{n-1} | x_{n-2}, \dots x_1) \dots P(x_2|x_1) P(x_1)$$

eg $P(x,y,z) = P(x|y,z)P(y|z)P(z)$

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

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

$\vdots$

etc.

Product decomposition

Combining the chain rule & the conditional independences implied by a DAG gives, for a set of N random variables $X_1, \dots, X_N$ gives :

$$P(x_1, x_2, \dots, x_N) = \prod_{i=1}^{N} P(x_i | pa(x_i))$$

Notes:

- this does not rely on $X_1, X_2, \dots X_N$ being 'in order' according to the DAG
  - → this is why there is the '$pa(x_i)$' in the expression
- In general this allows us to specify the joint distribution as a product of simpler distributions similar to the case of mutual independence

Other independencies in DAGs

A given DAG tells us 'directly' about the dependencies of a node on its ancestors ('past' nodes)

→ it doesn't 'directly' tell us about the (probabilistic) dependencies of a node on its descendents!

eg $X \to Y$

tells us Y depends on X

& $P(x,y) = P(y|x)P(x)$

↑ $Pa(y)$ ↑ no parents

But $P(x,y) = P(x|y)P(y)$ always] prob. theory

& $P(x|y) = \frac{P(y|x)P(x)}{P(y)} \neq P(x)$

$\Rightarrow$ X is not indep. of Y in $X \to Y$

Causal vs probabilistic dependence?

- The DAG interpretation so far is purely probabilistic, not 'causal'
- We can add interpretational components so that $X \to Y$ also captures 'X causes Y'...

BUT for now just focus on the probabilistic conditional (in)dependencies implied by a DAG

→ combine ordered Markov factorisation & probability theory

--- OR use graph theory!

Further conditional independences implied by a DAG: directed separation

To determine the other independences implied by the directed Markov factorisation (encoded by the DAG) & the rules of probability, we can consider the graphical concept of

'd-separation'

(for directed separation)

→ For this we will call any sequence of edges between two nodes, regardless of direction, a dependence path or just 'path':

$A \to B \leftarrow C$ } $A \to B, B \leftarrow C$ is a path

d-separation

[More complex than u-separation unfortunately!]

Definition: **path blocking**

A (dependence) path between any two nodes X, Y in a DAG is 'blocked' by a set of nodes $\mathcal{B}$ iff the path contains at least one sub-path ('junction') of the form:

1. $A \to B \to C$ (chain) where $B \in \mathcal{B}$

or 2. $A \leftarrow B \to C$ (fork) where $B \in \mathcal{B}$

or 3. $A \to D \leftarrow C$ (collider) where $D \notin \mathcal{B}$ & $\text{des}(D) \notin \mathcal{B}$

↑ descendents of D

d-separation:

> If a set of nodes $\mathcal{B}$ blocks every path between two nodes X & Y then X & Y are said to be 'd-separated given $\mathcal{B}$'

→ According to the standard 'directed Markov' interpretation of a DAG & standard probability theory, we have:

$$X \perp\!\!\!\perp Y \mid \mathcal{B}$$

$$\Leftrightarrow$$

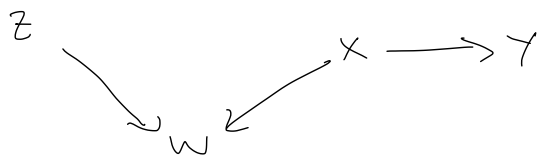
X & Y d-separated given $\mathcal{B}$

→ Furthermore:

> d-separation can derive all the conditional independences implied by the DAG!

Example

Consider:



- what can we say about the (in)dependence of Z & Y?

→ we can 'condition' on the empty set to get 'unconditional' independence properties

→ Note that the dependence path between Z & Y contains the sub path $Z \rightarrow W \leftarrow X$, ie a 'collider'

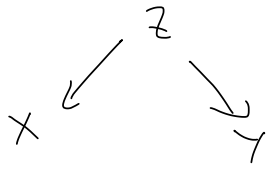
→ Hence $Z \perp\!\!\!\perp Y \mid \{\} \Leftrightarrow Z \perp\!\!\!\perp Y$

But $Z \not\perp\!\!\!\perp Y \mid W$ 'unblocks' path

Note: If two nodes are not d-separated, they are called d-connected & are (likely) 'not independent' ie '$X \not\perp\!\!\!\perp Y$'

Exercises

1. Consider



which of these is true:

$X \perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y \mid Z$

2. Suppose in (1) we are sampling from a population & measuring variables:

Z = 'age'

X = 'retirement savings'

Y = 'number of wrinkles'

What can you say about the relationship between retirement savings & wrinkles?

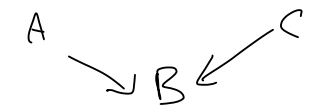
How should you analyse this relationship to understand the 'causal effects'?

3. Consider $A \to B \to C$, $A \leftarrow B \to C$

& $A \leftarrow B \leftarrow C$

- Derive the implied conditional independencies using d-separation
- what do you notice?

4. Consider:



Is $A \perp\!\!\!\perp C$?

Is $A \perp\!\!\!\perp C \mid B$?

Suppose A represents 'height'

B represents 'basketball success'

C represents 'basketball shooting skill'

in the above.

Interpret the associated conditional independencies.

5. Further reading:

d-separation & associated independence properties can be extended to sets of variables

→ see Wasserman (2004) p. 270 (attached)

6. Further reading:

Look up 'Simpson's paradox' & its relation to DAGs.

7. Further reading:

'd-separation without tears' at:

http://bayes.cs.ucla.edu/BOOK-2K/d-sep.html

Wasserman (2004) 'All of statistics'

17

Directed Graphs and Conditional Independence

17.1 Introduction

A directed graph consists of a set of nodes with arrows between some nodes. An example is shown in Figure 17.1.

Graphs are useful for representing independence relations between variables. They can also be used as an alternative to counterfactuals to represent causal relationships. Some people use the phrase **Bayesian network** to refer to a directed graph endowed with a probability distribution. This is a poor choice of terminology. Statistical inference for directed graphs can be performed using

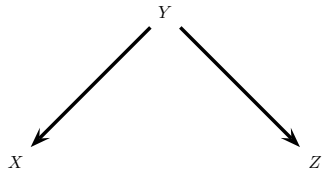


FIGURE 17.1. A directed graph with vertices $V = \{X, Y, Z\}$ and edges $E = \{(Y, X), (Y, Z)\}$.

frequentist or Bayesian methods, so it is misleading to call them Bayesian networks.

Before getting into details about directed acyclic graphs (DAGs), we need to discuss conditional independence.

17.2 Conditional Independence

17.1 Definition. *Let X, Y and Z be random variables. X and Y are* **conditionally independent given** *Z, written $X \amalg Y \mid Z$, if*

$$f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z). \tag{17.1}$$

for all x, y and z.

Intuitively, this means that, once you know Z, Y provides no extra information about X. An equivalent definition is that

$$f(x|y,z) = f(x|z). \tag{17.2}$$

The conditional independence relation satisfies some basic properties.

17.2 Theorem. *The following implications hold:* [1]

$$\begin{aligned}
X \amalg Y \mid Z &\implies Y \amalg X \mid Z \\
X \amalg Y \mid Z \text{ and } U = h(X) &\implies U \amalg Y \mid Z \\
X \amalg Y \mid Z \text{ and } U = h(X) &\implies X \amalg Y \mid (Z,U) \\
X \amalg Y \mid Z \text{ and } X \amalg W \mid (Y,Z) &\implies X \amalg (W,Y) \mid Z \\
X \amalg Y \mid Z \text{ and } X \amalg Z \mid Y &\implies X \amalg (Y,Z).
\end{aligned}$$

17.3 DAGs

A **directed graph** $\mathcal{G}$ consists of a set of vertices V and an edge set E of ordered pairs of vertices. For our purposes, each vertex will correspond to a random variable. If $(X,Y) \in E$ then there is an arrow pointing from X to Y. See Figure 17.1.

[1] The last property requires the assumption that all events have positive probability; the first four do not.

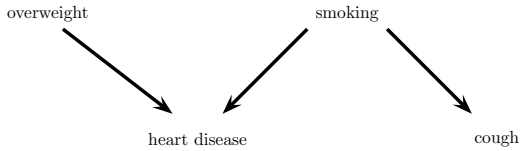
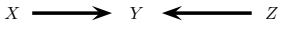


FIGURE 17.2. DAG for Example 17.4.

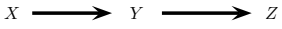
If an arrow connects two variables X and Y (in either direction) we say that X and Y are **adjacent.** If there is an arrow from X to Y then X is a **parent** of Y and Y is a **child** of X. The set of all parents of X is denoted by π_X or $\pi(X)$. A **directed path** between two variables is a set of arrows all pointing in the same direction linking one variable to the other such as:

A sequence of adjacent vertices staring with X and ending with Y but ignoring the direction of the arrows is called an **undirected path.** The sequence $\{X,Y,Z\}$ in Figure 17.1 is an undirected path. X is an **ancestor** of Y if there is a directed path from X to Y (or $X = Y$). We also say that Y is a **descendant** of X.

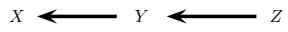
A configuration of the form:



is called a **collider** at Y. A configuration not of that form is called a **non-collider**, for example,



or



The collider property is path dependent. In Figure 17.7, Y is a collider on the path $\{X, Y, Z\}$ but it is a non-collider on the path $\{X, Y, W\}$. When the variables pointing into the collider are not adjacent, we say that the collider is **unshielded**. A directed path that starts and ends at the same variable is called a **cycle.** A directed graph is **acyclic** if it has no cycles. In this case we say that the graph is a **directed acyclic graph** or **DAG**. From now on, we only deal with acyclic graphs.

17.4 Probability and DAGs

Let $\mathcal{G}$ be a DAG with vertices $V = (X_1, \ldots, X_k)$.

17.3 Definition. *If $\mathbb{P}$ is a distribution for V with probability function f, we say that $\mathbb{P}$ is* **Markov to** $\mathcal{G}$*, or that $\mathcal{G}$* **represents** *$\mathbb{P}$, if*

$$f(v) = \prod_{i=1}^{k} f(x_i \mid \pi_i) \tag{17.3}$$

where π_i are the parents of X_i. The set of distributions represented by $\mathcal{G}$ is denoted by $M(\mathcal{G})$.

17.4 Example. Figure 17.2 shows a DAG with four variables. The probability function for this example factors as

$$\begin{aligned} & f(\text{overweight}, \text{smoking}, \text{heart disease}, \text{cough}) \\ &\quad = f(\text{overweight}) \times f(\text{smoking}) \\ &\quad \times f(\text{heart disease} \,|\, \text{overweight}, \text{smoking}) \\ &\quad \times f(\text{cough} \,|\, \text{smoking}). \quad \blacksquare \end{aligned}$$

17.5 Example. For the DAG in Figure 17.3, $\mathbb{P} \in M(\mathcal{G})$ if and only if its probability function f has the form

$$f(x, y, z, w) = f(x) f(y) f(z \mid x, y) f(w \mid z). \quad \blacksquare$$

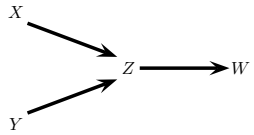


FIGURE 17.3. Another DAG.

The following theorem says that $\mathbb{P} \in M(\mathcal{G})$ if and only if the **Markov Condition** holds. Roughly speaking, the Markov Condition means that every variable W is independent of the "past" given its parents.

17.6 Theorem. *A distribution $\mathbb{P} \in M(\mathcal{G})$ if and only if the following* **Markov Condition** *holds: for every variable W,*

$$W \amalg \widetilde{W} \mid \pi_W \tag{17.4}$$

where $\widetilde{W}$ denotes all the other variables except the parents and descendants of W.

17.7 Example. In Figure 17.3, the Markov Condition implies that

$$X \amalg Y \quad \text{and} \quad W \amalg \{X, Y\} \mid Z. \quad \blacksquare$$

17.8 Example. Consider the DAG in Figure 17.4. In this case probability function must factor like

$$f(a, b, c, d, e) = f(a) f(b|a) f(c|a) f(d|b, c) f(e|d).$$

The Markov Condition implies the following independence relations:

$$D \amalg A \mid \{B, C\}, \quad E \amalg \{A, B, C\} \mid D \quad \text{and} \quad B \amalg C \mid A \quad \blacksquare$$

17.5 More Independence Relations

The Markov Condition allows us to list some independence relations implied by a DAG. These relations might imply other independence relations. Con-

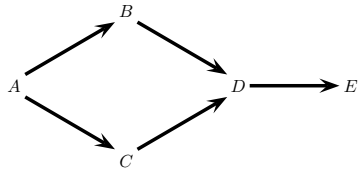


FIGURE 17.4. Yet another DAG.

sider the DAG in Figure 17.5. The Markov Condition implies:

$$X_1 \amalg X_2, \qquad X_2 \amalg \{X_1, X_4\}, \qquad X_3 \amalg X_4 \mid \{X_1, X_2\},$$

$$X_4 \amalg \{X_2, X_3\} \mid X_1, \qquad X_5 \amalg \{X_1, X_2\} \mid \{X_3, X_4\}$$

It turns out (but it is not obvious) that these conditions imply that

$$\{X_4, X_5\} \amalg X_2 \mid \{X_1, X_3\}.$$

How do we find these extra independence relations? The answer is "d-separation" which means "directed separation." d-separation can be summarized by three rules. Consider the four DAG's in Figure 17.6 and the DAG in Figure 17.7. The first 3 DAG's in Figure 17.6 have no colliders. The DAG in the lower right of Figure 17.6 has a collider. The DAG in Figure 17.7 has a collider with a descendant.

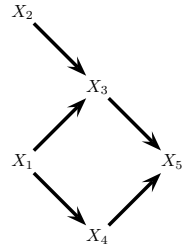


FIGURE 17.5. And yet another DAG.

FIGURE 17.6. The first three DAG's have no colliders. The fourth DAG in the lower right corner has a collider at Y.

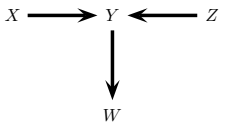


FIGURE 17.7. A collider with a descendant.

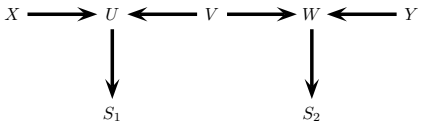


FIGURE 17.8. d-separation explained.

The Rules of d-Separation

Consider the DAGs in Figures 17.6 and 17.7.

1. When Y is not a collider, X and Z are **d-connected**, but they are **d-separated** given Y.
2. If X and Z collide at Y, then X and Z are **d-separated**, but they are **d-connected** given Y.
3. Conditioning on the descendant of a collider has the same effect as conditioning on the collider. Thus in Figure 17.7, X and Z are **d-separated** but they are **d-connected** given W.

Here is a more formal definition of d-separation. Let X and Y be distinct vertices and let W be a set of vertices not containing X or Y. Then X and Y are **d-separated given** W if there exists no undirected path U between X and Y such that (i) every collider on U has a descendant in W, and (ii) no other vertex on U is in W. If A, B, and W are distinct sets of vertices and A and B are not empty, then A and B are d-separated given W if for every $X \in A$ and $Y \in B$, X and Y are d-separated given W. Sets of vertices that are not d-separated are said to be d-connected.

17.9 Example. Consider the DAG in Figure 17.8. From the d-separation rules we conclude that:
X and Y are d-separated (given the empty set);
X and Y are d-connected given $\{S_1, S_2\}$;
X and Y are d-separated given $\{S_1, S_2, V\}$.

17.10 Theorem. [2] *Let A, B, and C be disjoint sets of vertices. Then $A \amalg B \mid C$ if and only if A and B are d-separated by C.*

[2] We implicitly assume that $\mathbb{P}$ is **faithful** to $\mathcal{G}$ which means that $\mathbb{P}$ has no extra independence relations other than those logically implied by the Markov Condition.

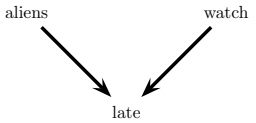


FIGURE 17.9. Jordan's alien example (Example 17.11). Was your friend kidnapped by aliens or did you forget to set your watch?

17.11 Example. The fact that conditioning on a collider creates dependence might not seem intuitive. Here is a whimsical example from Jordan (2004) that makes this idea more palatable. Your friend appears to be late for a meeting with you. There are two explanations: she was abducted by aliens or you forgot to set your watch ahead one hour for daylight savings time. (See Figure 17.9.) Aliens and Watch are blocked by a collider which implies they are marginally independent. This seems reasonable since — before we know anything about your friend being late — we would expect these variables to be independent. We would also expect that $\mathbb{P}(\text{Aliens} = \text{yes} | \text{Late} = \text{yes}) > \mathbb{P}(\text{Aliens} = \text{yes})$; learning that your friend is late certainly increases the probability that she was abducted. But when we learn that you forgot to set your watch properly, we would lower the chance that your friend was abducted. Hence, $\mathbb{P}(\text{Aliens} = \text{yes} | \text{Late} = \text{yes}) \neq \mathbb{P}(\text{Aliens} = \text{yes} | \text{Late} = \text{yes}, \text{Watch} = \text{no})$. Thus, Aliens and Watch are dependent given Late. ■

17.12 Example. Consider the DAG in Figure 17.2. In this example, overweight and smoking are marginally independent but they are dependent given heart disease. ■

Graphs that look different may actually imply the same independence relations. If $\mathcal{G}$ is a DAG, we let $\mathcal{I}(\mathcal{G})$ denote all the independence statements implied by $\mathcal{G}$. Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ for the same variables V are **Markov equivalent** if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$. Given a DAG $\mathcal{G}$, let skeleton($\mathcal{G}$) denote the undirected graph obtained by replacing the arrows with undirected edges.

17.13 Theorem. *Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if (i)* skeleton($\mathcal{G}_1$) = skeleton($\mathcal{G}_2$) *and (ii) $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same unshielded colliders.*

17.14 Example. The first three DAGs in Figure 17.6 are Markov equivalent. The DAG in the lower right of the Figure is not Markov equivalent to the others. ■