Bayesuvius,
a small visual dictionary of Bayesian Networks

Robert R. Tucci
www.ar-tiste.xyz

October 1, 2020

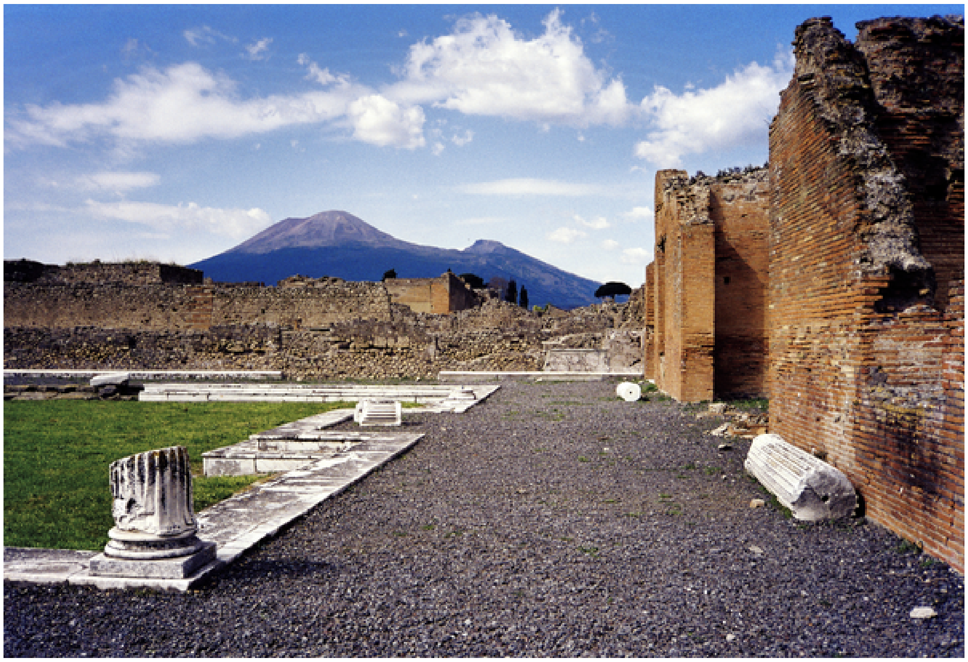


Figure 1: View of Mount Vesuvius from Pompeii

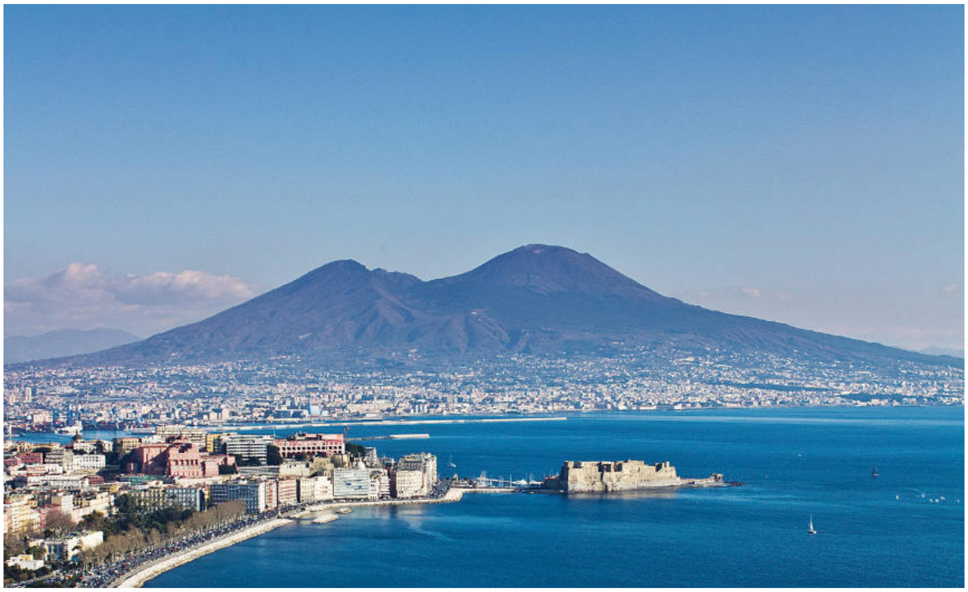


Figure 2: Mount Vesuvius and Bay of Naples

Contents

0.2 Definition of a Bayesian Network

A Bayesian network (bnet) consists of a DAG (Directed Acyclic Graph) and a Transition Probability Matrix (TPM) associated with each node of the graph. A TPM is often called a Conditional Probability Table (CPT).

In this book, random variables are indicated by underlined letters and their values by non-underlined letters. Each node of a bnet is labelled by a random variable. Thus, $\underline{x} = x$ means that node $\underline{x}$ is in state x.

Some sets of nodes associated with each node $\underline{a}$ of a bnet

- $ch(\underline{a})$ = children of $\underline{a}$.
- $pa(\underline{a})$ = parents of $\underline{a}$.
- $nb(\underline{a}) = pa(\underline{a}) \cup ch(\underline{a})$ = neighbors of $\underline{a}$.
- $de(\underline{a}) = \cup_{n=1}^{\infty} ch^n(\underline{a}) = ch(\underline{a}) \cup ch \circ ch(\underline{a}) \cup \ldots$, descendants of $\underline{a}$.
- $an(\underline{a}) = \cup_{n=1}^{\infty} pa^n(\underline{a}) = pa(\underline{a}) \cup pa \circ pa(\underline{a}) \cup \ldots$, ancestors of $\underline{a}$.

In this book, we will use $\underline{a}.$ to indicate a **multi-node (node set, node array)** $\underline{a}. = (\underline{a}_j)_{j=0,1,\ldots,na-1}$. We will often treat multinodes as if they were sets, and combine them with the usual set operators. For instance, for two multinodes $\underline{a}.$ and $\underline{b}.$, we define $\underline{a}. \cup \underline{b}.$, $\underline{a}. \cap \underline{b}.$, $\underline{a}. - \underline{b}.$ and $\underline{a}. \subset \underline{b}.$ in the obvious way. We will indicate a singleton set (single node multi-node) $\underline{a}. = \{\underline{a}\}$ simply by $\underline{a}. = \underline{a}$. For instance, $\underline{a}. - \underline{b} = \underline{a}. - \{\underline{b}\}$.

The TPM of a node $\underline{x}$ of a bnet is a matrix of probabilities $P(\underline{x} = x | pa(\underline{x}) = a.)$.

A bnet with nodes $\underline{x}.$ represents a probability distribution

$$P(x.) = \prod_j P(\underline{x}_j = x_j | pa(\underline{x}_j) = pa(x_j)) \ . \tag{1}$$

0.3 Notational Conventions and Preliminaries

Some abbreviations frequently used throughout this book.

- bnet= B net= Bayesian Network
- CPT = Conditional Probabilities Table, same as TPM
- DAG = Directed Acyclic Graph
- i.i.d.= independent identically distributed.
- TPM= Transition Probability Matrix, same as CPT

Define $\mathbb{Z}, \mathbb{R}, \mathbb{C}$ to be the integers, real numbers and complex numbers, respectively.

For $a < b$, define $\mathbb{Z}_I$ to be the integers in the interval I, where $I = [a, b], [a, b), (a, b], (a, b)$ (i.e, I can be closed or open on either side).

$A_{>0} = \{k \in A : k > 0\}$ for $A = \mathbb{Z}, \mathbb{R}$.

Random variables will be indicated by underlined letters and their values by non-underlined letters. Each node of a bnet will be labelled by a random variable. Thus, $\underline{x} = x$ means that node $\underline{x}$ is in state x.

It is more conventional to use an upper case letter to indicate a random variable and a lower case letter for its state. Thus, $X = x$ means that random variable X is in state x. However, we have opted in this book to avoid that notation, because we often want to define certain lower case letters to be random variables or, conversely, define certain upper case letters to be non-random variables.

$P_{\underline{x}}(x) = P(\underline{x} = x) = P(x)$ is the probability that random variable $\underline{x}$ equals $x \in S_{\underline{x}}$. $S_{\underline{x}}$ is the set of states (i.e., values) that $\underline{x}$ can assume and $n_{\underline{x}} = |S_{\underline{x}}|$ is the size (aka cardinality) of that set. Hence,

$$\sum_{x \in S_{\underline{x}}} P_{\underline{x}}(x) = 1 \tag{2}$$

$$P_{\underline{x},\underline{y}}(x, y) = P(\underline{x} = x, \underline{y} = y) = P(x, y) \tag{3}$$

$$P_{\underline{x}|\underline{y}}(x|y) = P(\underline{x} = x|\underline{y} = y) = P(x|y) = \frac{P(x, y)}{P(y)} \tag{4}$$

Kronecker delta function: For x, y in discrete set S,

$$\delta(x, y) = \begin{cases} 1 \text{ if } x = y \\ 0 \text{ if } x \neq y \end{cases} \tag{5}$$

Dirac delta function: For $x, y \in \mathbb{R}$,

$$\int_{-\infty}^{+\infty} dx\ \delta(x - y) f(x) = f(y) \tag{6}$$

The TPM of a node of a bnet can be either a discrete or a continuous probability distribution. To go from continuous to discrete, one replaces integrals over states of a node by sums over new states, and Dirac delta functions by Kronecker delta functions. More precisely, consider a function $f : [a, b] \to \mathbb{R}$. Express $[a, b]$ as a union of small, disjoint (except for one point) closed sub-intervals (bins) of length Δx. Name one point in each bin to be the representative of that bin, and let $S_{\underline{x}}$ be the set of all the bin representatives. This is called discretization or binning. Then

$$\frac{1}{(b-a)}\int_{[a,b]} dx\; f(x) \to \frac{\Delta x}{(b-a)}\sum_{x\in S_{\underline{x}}} f(x) = \frac{1}{n_{\underline{x}}}\sum_{x\in S_{\underline{x}}} f(x)\ . \tag{7}$$

Both sides of last equation are 1 when $f(x) = 1$. Furthermore, if $y \in S_{\underline{x}}$, then

$$\int_{[a,b]} dx\; \delta(x-y)f(x) = f(y) \to \sum_{x\in S_{\underline{x}}} \delta(x,y)f(x) = f(y)\ . \tag{8}$$

Indicator function (aka Truth function):

$$\mathbb{1}(\mathcal{S}) = \left\{\begin{array}{l} 1 \text{ if } \mathcal{S} \text{ is true} \\ 0 \text{ if } \mathcal{S} \text{ is false}\end{array}\right. \tag{9}$$

For example, $\delta(x, y) = \mathbb{1}(x = y)$.

$$\vec{x} = (x[0], x[1], x[2]\ldots, x[nsam(\vec{x}) - 1]) = x[:] \tag{10}$$

$nsam(\vec{x})$ is the number of samples of $\vec{x}$. $\underline{x}[i] \in S_{\underline{x}}$ are i.i.d. (independent identically distributed) samples with

$$x[i] \sim P_{\underline{x}}\ \text{ (i.e. } P_{\underline{x[i]}} = P_{\underline{x}}) \tag{11}$$

$$P(\underline{x} = x) = \frac{1}{nsam(\vec{x})}\sum_i \mathbb{1}(x[i] = x) \tag{12}$$

Hence, for any $f : S_{\underline{x}} \to \mathbb{R}$,

$$\sum_x P(\underline{x} = x)f(x) = \frac{1}{nsam(\vec{x})}\sum_i f(x[i]) \tag{13}$$

If we use two sampled variables, say $\vec{x}$ and $\vec{y}$, in a given bnet, their number of samples $nsam(\vec{x})$ and $nsam(\vec{y})$ need not be equal.

$$P(\vec{x}) = \prod_i P(x[i]) \tag{14}$$

$$\sum_{\vec{x}} = \prod_i \sum_{x[i]} \tag{15}$$

$$\partial_{\vec{x}} = [\partial_{x[0]}, \partial_{x[1]}, \partial_{x[2]}, \ldots, \partial_{x[nsam(\vec{x})-1]}] \tag{16}$$

$$P(\vec{x}) \approx [\prod_x P(x)^{P(x)}]^{nsam(\vec{x})} \tag{17}$$

$$= e^{nsam(\vec{x})\sum_x P(x)\ln P(x)} \tag{18}$$

$$= e^{-nsam(\vec{x})H(P_{\underline{x}})} \tag{19}$$

$$f^{[1,\partial_x,\partial_y]}(x,y) = [f, \partial_x f, \partial_y f] \tag{20}$$

$$f^+ = f^{[1,\partial_x,\partial_y]} \tag{21}$$

For probabilty distributions $p(x), q(x)$ of $x \in S_{\underline{x}}$

- Entropy:
$$H(p) = -\sum_x p(x)\ln p(x) \geq 0 \tag{22}$$

- Kullback-Liebler divergence:
$$D_{KL}(p \parallel q) = \sum_x p(x)\ln\frac{p(x)}{q(x)} \geq 0 \tag{23}$$

- Cross entropy:
$$CE(p \rightarrow q) = -\sum_x p(x)\ln q(x) \tag{24}$$
$$= H(p) + D_{KL}(p \parallel q) \tag{25}$$

Normal Distribution: $x, \mu, \sigma \in \mathbb{R}$, $\sigma > 0$

$$\mathcal{N}(\mu,\sigma^2)(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{26}$$

Uniform Distribution: $a < b$, $x \in [a,b]$

$$\mathcal{U}(a,b)(x) = \frac{1}{b-a} \tag{27}$$

Expected Value and Variance

Given a random variable $\underline{x}$ with states $S_{\underline{x}}$ and a function $f : S_{\underline{x}} \rightarrow \mathbb{R}$, define

$$E_{\underline{x}}[f(\underline{x})] = E_{x\sim P(x)}[f(x)] = \sum_x P(x)f(x) \tag{28}$$

$$Var_{\underline{x}}[f(\underline{x})] = E_{\underline{x}}\left[(f(\underline{x}) - E_{\underline{x}}[f(\underline{x})])^2\right] \tag{29}$$

$$= E_{\underline{x}}[f(\underline{x})^2] - (E_{\underline{x}}[f(\underline{x})])^2 \tag{30}$$

$$E[\underline{x}] = E_{\underline{x}}[\underline{x}] \tag{31}$$

$$Var[\underline{x}] = Var_{\underline{x}}[\underline{x}] \tag{32}$$

Conditional Expected Value

Given a random variable $\underline{x}$ with states $S_{\underline{x}}$, a random variable $\underline{y}$ with states $S_{\underline{y}}$, and a function $f : S_{\underline{x}} \times S_{\underline{y}} \to \mathbb{R}$, define

$$E_{\underline{x}|\underline{y}}[f(\underline{x},\underline{y})] = \sum_x P(x|\underline{y}) f(x,\underline{y}) \ , \tag{33}$$

$$E_{\underline{x}|\underline{y}=y}[f(\underline{x},y)] = E_{\underline{x}|y}[f(\underline{x},y)] = \sum_x P(x|y) f(x,y) \ . \tag{34}$$

Note that

$$\begin{aligned} E_{\underline{y}}[E_{\underline{x}|\underline{y}}[f(\underline{x},\underline{y})]] &= \sum_{x,y} P(x|y)P(y)f(x,y) \qquad (35) \\ &= \sum_{x,y} P(x,y)f(x,y) \qquad (36) \\ &= E_{\underline{x},\underline{y}}[f(\underline{x},\underline{y})] \ . \qquad (37) \end{aligned}$$

Law of Total Variance

Claim 1 *Suppose $P : S_{\underline{x}} \times S_{\underline{y}} \to [0,1]$ is a probability distribution. Suppose $f : S_{\underline{x}} \times S_{\underline{y}} \to \mathbb{R}$ and $f = f(x,y)$. Then*

$$Var_{\underline{x},\underline{y}}(f) = E_{\underline{y}}[Var_{\underline{x}|\underline{y}}(f)] + Var_{\underline{y}}(E_{\underline{x}|\underline{y}}[f]) \ . \tag{38}$$

In particular,

$$Var_{\underline{x}}(x) = E_{\underline{y}}[Var_{\underline{x}|\underline{y}}(x)] + Var_{\underline{y}}(E_{\underline{x}|\underline{y}}[x]) \ . \tag{39}$$

proof:

Let

$$A = \sum_y P(y) \left(\sum_x P(x|y) f \right)^2 \ . \tag{40}$$

Then

$$\begin{aligned} Var_{\underline{x},\underline{y}}(f) &= \sum_{x,y} P(x,y)f^2 - \left(\sum_{x,y} P(x,y) f \right)^2 \qquad (41) \\ &= \left\{ \begin{array}{l} \sum_{x,y} P(x,y)f^2 - A \\ + \left(A - \left(\sum_{x,y} P(x,y) f \right)^2 \right) \end{array} \right. \qquad (42) \end{aligned}$$

$$
\begin{aligned}
E_{\underline{y}}[Var_{\underline{x}|\underline{y}}(f)] &= \sum_y P(y)\left(\sum_x P(x|y)f^2 - \left(\sum_x P(x|y)f\right)^2\right) && (43)\\
&= \sum_{x,y} P(x,y)f^2 - A && (44)
\end{aligned}
$$

$$
\begin{aligned}
Var_{\underline{y}}(E_{\underline{x}|\underline{y}}[f]) &= \sum_y P(y)\left(\sum_x P(x|y)f\right)^2 - \left(\sum_y P(y)\sum_x P(x|y)f\right)^2 && (45)\\
&= A - \left(\sum_{x,y} P(x,y)f\right)^2 && (46)
\end{aligned}
$$

QED

$\langle\underline{x},\underline{y}\rangle$ notation, for covariances of any two random variables $\underline{x},\underline{y}$.

Mean value of $\underline{x}$

$$\langle\underline{x}\rangle = E_{\underline{x}}[\underline{x}] \tag{47}$$

Signed distance of $\underline{x}$ to its mean value

$$\Delta\underline{x} = \underline{x} - \langle\underline{x}\rangle \tag{48}$$

Covariance of $(\underline{x},\underline{y})$

$$\langle\underline{x},\underline{y}\rangle = \langle\Delta\underline{x}\Delta\underline{y}\rangle = Cov(\underline{x},\underline{y}) \tag{49}$$

Variance of $\underline{x}$

$$Var(\underline{x}) = \langle\underline{x},\underline{x}\rangle \tag{50}$$

Standard deviation or $\underline{x}$

$$\sigma_{\underline{x}} = \sqrt{\langle\underline{x},\underline{x}\rangle} \tag{51}$$

Correlation of $(\underline{x},\underline{y})$

$$\rho_{\underline{x},\underline{y}} = \frac{\langle\underline{x},\underline{y}\rangle}{\sqrt{\langle\underline{x},\underline{x}\rangle\langle\underline{y},\underline{y}\rangle}} \tag{52}$$

Sigmoid function: For $x \in \mathbb{R}$,

$$\text{sig}(x) = \frac{1}{1+e^{-x}} \tag{53}$$

$\mathcal{N}(!a)$ will denote a normalization constant that does not depend on a. For example, $P(x) = \mathcal{N}(!x)e^{-x}$ where $\int_0^\infty dx\, P(x) = 1$.

A **one hot** vector of zeros and ones is a vector with all entries zero with the exception of a single entry which is one. A **one cold** vector has all entries equal to one with the exception of a single entry which is zero. For example, if $x^n = (x_0, x_1, \ldots, x_{n-1})$ and $x_i = \delta(i, 0)$ then x^n is one hot.

Short Summary of Boolean Algebra.
See Ref.[5] for more info about this topic.

Suppose $x, y, z \in \{0, 1\}$. Define

$$x \text{ or } y = x \vee y = x + y - xy \ , \tag{54}$$

$$x \text{ and } y = x \wedge y = xy \ , \tag{55}$$

and

$$\text{not } x = \overline{x} = 1 - x \ , \tag{56}$$

where we are using normal addition and multiplication on the right hand sides.[1]

Associativity	$x \vee (y \vee z) = (x \vee y) \vee z$ $x \wedge (y \wedge z) = (x \wedge y) \wedge z$
Commutativity	$x \vee y = y \vee x$ $x \wedge y = y \wedge x$
Distributivity	$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$
Identity	$x \vee 0 = x$ $x \wedge 1 = x$
Annihilator	$x \wedge 0 = 0$ $x \vee 1 = 1$
Idempotence	$x \vee x = x$ $x \wedge x = x$
Absorption	$x \wedge (x \vee y) = x$ $x \vee (x \wedge y) = x$
Complementation	$x \wedge \overline{x} = 0$ $x \vee \overline{x} = 1$
Double negation	$\overline{(\overline{x})} = x$
De Morgan Laws	$\overline{x} \wedge \overline{y} = \overline{(x \vee y)}$ $\overline{x} \vee \overline{y} = \overline{(x \wedge y)}$

Table 1: Boolean Algebra Identities

Actually, since $x \wedge y = xy$, we can omit writing the symbol $\wedge$. The symbol $\wedge$ is useful to exhibit the symmetry of the identities, and to remark about the analogous identities for sets, where $\wedge$ becomes intersection $\cap$ and $\vee$ becomes union $\cup$. However, for practical calculations, $\wedge$ is an unnecessary nuisance.

Since $x \in \{0, 1\}$,

$$P(\overline{x}) = 1 - P(x) \ . \tag{57}$$

[1] Note the difference between $\vee$ and modulus 2 addition $\oplus$. For $\oplus$ (aka XOR): $x \oplus y = x + y - 2xy$.

Clearly, from analyzing the simple event space $(x, y) \in \{0, 1\}^2$,

$$P(x \vee y) = P(x) + P(y) - P(x \wedge y) \ . \tag{58}$$

0.4 Navigating the ocean of Judea Pearl's Books

Many of the greatest ideas in the bnet field were invented by Judea Pearl and his collaborators. Thus, this book is heavily indebted to those great scientists. Those ideas have had no clearer and more generous expositor than Judea Pearl himself.

Pearl has written 4 books that I have used in writing Bayesuvius. His 1988 book Ref.[1] dates back to his pre-causal period. That book I used to learn about topics such as d-separation, belief propagation, Markov-blankets, and noisy-ORs. 3 other books that he wrote later, in his causal period, are:

1. In 2000 (1st ed.), and 2013 (2nd ed.), Pearl published what is so far his most technical and exhaustive book on the subject of causality, Ref[2].

2. In 2016, he released together with Glymour and Jewell, a less advanced "primer" on causality, Ref.[3].

3. In 2018, he released together with Mackenzie his lovely "The Book of Why", Ref.[4].

Those 3 books I used to learn about causality topics such as do-calculus, backdoor and front-door adjustments, linear deterministic bnets with exogenous noise, and counterfactuals.

Chapter 1

Backdoor Adjustment

The backdoor (BD) adjustment theorem is proven in Chapter 9 from the rules of do-calculus. The goal of this chapter is to give examples of the use of that theorem. We will restate the theorem in this chapter, sans proof. There is no need to understand the theorem's proof in order to use it. However, you will need to skim Chapter 9 in order to familiarize yourself with the notation used to state the theorem. This chapter also assumes that you are comfortable with the rules for checking for d-separation. Those rules are covered in Chapter 10.

Suppose we have access to data that allows us to estimate a probability distribution $P(x., y., z.)$. Hence, the variables $\underline{x}., \underline{y}., \underline{z}.$ are all observed (i.e, not hidden). Then we say that the backdoor $\underline{z}.$ satisfies the **backdoor criterion** relative to $(\underline{x}., \underline{y}.)$ if

1. All paths from $\underline{x}.$ to $\underline{y}.$ that start with an arrow pointing into $\underline{x}.$, are blocked by $\underline{z}.$.
2. $\underline{z}. \notin de(\underline{x}.)$.

Claim 2 Backdoor Adjustment Theorem

If $\underline{z}.$ satisfies the backdoor criterion relative to $(\underline{x}., \underline{y}.)$, then

$$P(y.|\rho\underline{x}. = x.) = \sum_{z.} P(y.|x., z.)P(z.) \tag{1.1}$$

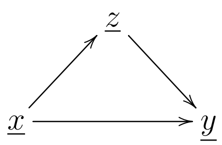
$$= \sum_{z.} \left\{ \begin{tikzcd} \underline{z}. = z. \arrow[dr] & \\ \underline{x}. = x. \arrow[r] & \underline{y}. \end{tikzcd} \right\} \tag{1.2}$$

proof: See Chapter 9

QED

Examples:

1.

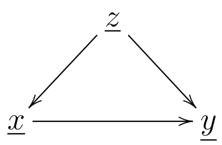


(1.3)

BD criterion satisfied if $\underline{x}. = \underline{x}, \underline{y}. = \underline{y}, \underline{z}. = \emptyset$. No adjustment necessary.

$$P(y|\rho\underline{x} = x) = P(y|x) \tag{1.4}$$

2.

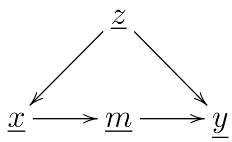


(1.5)

BD criterion satisfied if $\underline{x}. = \underline{x}, \underline{y}. = \underline{y}, \underline{z}. = \underline{z}$.

Note that here the backdoor formula adjusts the parents of $\underline{x}$..

3.

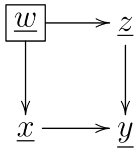


(1.6)

BD criterion satisfied if $\underline{x}. = \underline{x}, \underline{y}. = \underline{y}, \underline{z}. = \underline{z}$.

This bnet is also used to demonstrate the front-door criterion.

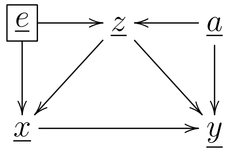
4.



(1.7)

BD criterion satisfied if $\underline{x}. = \underline{x}, \underline{y}. = \underline{y}, \underline{z}. = \underline{z}$. Note that here the backdoor formula cannot adjust the single parent $\underline{w}$ of $\underline{x}$ because it is hidden, but we are able to block the backdoor path by conditioning on $\underline{z}$ instead.

5.

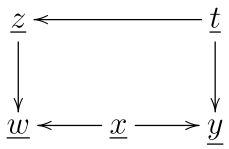


(1.8)

Conditioning on $\underline{z}$ blocks backdoor path $\underline{x} - \underline{z} - \underline{y}$, but opens path $\underline{x} - \underline{e} - \underline{z} - \underline{a} - \underline{y}$ because $\underline{z}$ is a collider for that path. That path is blocked if we also condition on $\underline{a}$, which is possible because $\underline{a}$ is observed. In conclusion, the BD criterion is satisfied if $\underline{x}. = \underline{x}$, $\underline{y}. = \underline{y}$ and $\underline{z}. = (\underline{z}, \underline{a})$.

Conditioning on the parents of $\underline{x}.$ is often enough to block all backdoor paths. However, sometimes some of the parents are unobserved and one must condition on other nodes that are not parents of $\underline{x}.$ in order to satisfy the BD criterion.

6.



(1.9)

No need to control anything because only possible backdoor path is blocked by collider $\underline{w}$. Hence,

$$P(y|\rho\underline{x} = x) = P(y|x) \ . \tag{1.10}$$

However, if for some reason we want to control $\underline{w}$, we can block the path by controlling $\underline{t}$ too. Thus, the BD criterion is satisfied if $\underline{x}. = \underline{x}$, $\underline{y}. = \underline{y}$ and $\underline{z}. = (\underline{w}, \underline{t})$. Therefore,

$$P(y|\rho\underline{x} = x) = \sum_{t,w} P(y|x,t,w)P(t,w) \ . \tag{1.11}$$

Alternatively, can condition on $\underline{w}$ a priori, and satisfy the BD criterion with $\underline{x}. = \underline{x}$, $\underline{y}. = \underline{y}$ and $\underline{z}. = \underline{t}$; thus,

$$P(y|\rho\underline{x} = x, w) = \sum_{t} P(y|x,t,w)P(t|w) \ . \tag{1.12}$$

Multiplying Eq.(1.12) by $P(w)$ and summing over w gives Eq.(1.11).

7. Discuss reasons why multiple possible sets $\underline{z}.$ that satisfy the BD criterion can be advantageous.

- Can evaluate $P(y.|\rho\underline{x}. = x.)$ multiple ways and compare the results. This is a test that the causal bnet is correct.
- Some $\underline{z}.$ might be easier or less expensive to get data on.

Chapter 9

Do-Calculus

The do-calculus and associated ideas were invented by Judea Pearl and collaborators. This chapter is based on Judea Pearl's books. (See 0.4).

When doing do-calculus, it is convenient to separate the nodes of a bnet into 2 types: **visible (observed)**, and **non-visible (not observed, hidden)**, depending on whether data describing the state of that node is available (visible) or not (non-visible). In this chapter, hidden nodes will be indicated in a bnet diagram by either: (1) enclosing their random variable in a box (as if it were a black box) or (2) making the arrows coming out of them dashed. Accordingly, the 3 diagrams in Fig.9.1 all mean the same thing. A **confounder node for $\underline{x} \rightarrow \underline{y}$** (such as node $\underline{c}$ in Fig.9.1) is a hidden node with arrows pointing from it to both $\underline{x}$ and $\underline{y}$.

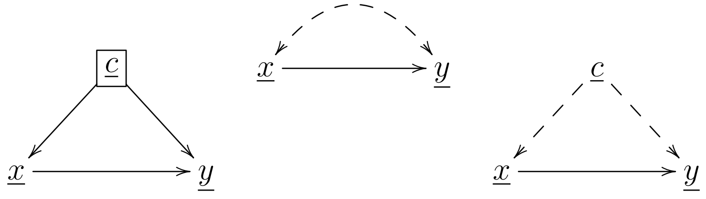


Figure 9.1: These 3 diagrams are equivalent. They mean that node $\underline{c}$ is hidden. Node $\underline{c}$ is implicit in the middle diagram.

Define an operator $\rho_{\underline{x}}$ that acts on a node $\underline{x}$ of a bnet to delete all the arrows entering $\underline{x}$, thus coverting $\underline{x}$ into a new node $\rho\underline{x}$ that is a root node. Define an analogous operator $\lambda\underline{x}$ that acts on a node $\underline{x}$ of a bnet to delete all the arrows leaving $\underline{x}$, thus coverting $\underline{x}$ into a new node $\lambda\underline{x}$ that is a leaf node. $\rho_{\underline{x}}$ and $\lambda_{\underline{x}}$ are depicted in Fig.9.2.

If you don't know yet what we mean by a a multi-node $\underline{a}.$, see Chapter 0.2

Given a bnet G, we define as follows the operators $\rho_{\underline{a}.}$ and $\lambda_{\underline{a}.}$ for a multi-node $\underline{a}.$.

$$\rho_{\underline{a}.}G = \left[\prod_j \rho_{\underline{a}_j}\right] G\ , \quad \lambda_{\underline{a}.}G = \left[\prod_j \lambda_{\underline{a}_j}\right] G\ . \tag{9.1}$$

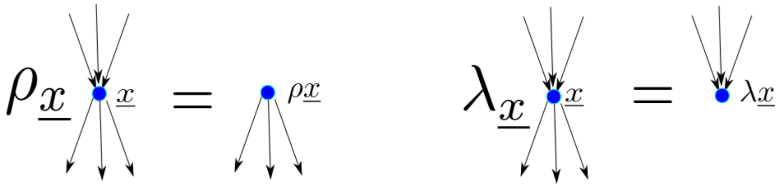


Figure 9.2: The operator $\rho_{\underline{x}}$ converts node $\underline{x}$ into a root node $\rho\underline{x}$. The operator $\lambda_{\underline{x}}$ converts node $\underline{x}$ into a leaf node $\lambda\underline{x}$.

Consider a bnet whose totality of nodes is labelled $\underline{X}.$. Recall that

$$P(X.) = \prod_j P(X_j|pa(X_j)) \; . \tag{9.2}$$

Now define an operator ρ that acts as follows[1]

$$P(X. - a.|\rho\underline{a}. = a.) = \mathcal{N}(!(X. - a.)) \prod_{j:\underline{X}_j \notin \underline{a}.} P(X_j|pa(X_j)) \; . \tag{9.3}$$

Furthermore, for $\underline{b}. \subset \underline{X}. - \underline{a}.$, define

$$P(b.|\rho\underline{a}. = a.) = \sum_{X.-a.-b.} P(X. - a.|\rho\underline{a}. = a.) \; , \tag{9.4}$$

and for $\underline{s}. \subset \underline{X}. - \underline{a}. - \underline{b}.$, define

$$P(b.|\rho\underline{a}. = a., s.) = \frac{P(b., s.|\rho\underline{a}. = a.)}{P(s.|\rho\underline{a}. = a.)} \; . \tag{9.5}$$

$P(b.|\rho\underline{a}. = a., s.)$ is usually denoted instead by $P(b.|do(\underline{a}. = a.), s.)$. I prefer to use ρ instead of $do()$ to remind me that it generates root nodes. I'll still call ρ a do operator.

$P(b.|\rho\underline{a}. = a., s.)$ is said to be **identifiable** if it can be expressed in terms of probability distributions that only depend on observed variables and that have no do operators in them.

For $\underline{x}, \underline{y} \in \{0, 1\}$, the "causal effect difference" , or "average causal effect" (ACE) is defined as

$$ACE = P(y = 1|\rho\underline{x} = 1) - P(y = 1|\rho\underline{x} = 0) \tag{9.6}$$

and the Risk Difference (RD) as

$$RD = P(y = 1|\underline{x} = 1) - P(y = 1|\underline{x} = 0) \; . \tag{9.7}$$

[1] As usual, $\mathcal{N}(!x)$ denotes a constant that is independent of x.

3 Rules of do-calculus

Throughout this section, suppose $\underline{a}., \underline{b}., \underline{r}., \underline{s}.$ are disjoint multinodes in a bnet G.

Recall from Chapter 10 on d-separation, that $(\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.)$ means that we have established from the d-separation rules that that all paths in G from $\underline{a}.$ to $\underline{b}.$ are blocked if we condition on $\underline{r}. \cup \underline{s}.$. Recall also that:

- **Rule 0:** Insertion or deletion of observations, without do operators. $(\underline{a}. = a. \leftrightarrow 1\ )$

  If $(\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.)$ in G, then $P(b.|a., r., s.) = P(b.|r., s.)$

The 3 rules of do-calculus can be presented in the same format.

- **Rule 1:** Insertion or deletion of observations $(\underline{a}. = a. \leftrightarrow 1\ )$

  If $(\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.)$ in $\rho_{\underline{r}.}G$, then $P(b.|a., \rho\underline{r}. = r., s.) = P(b.|\rho\underline{r}. = r., s.)$.

- **Rule 2:** Action or observation exchange $(\rho\underline{a}. = a. \leftrightarrow \underline{a}. = a.)$

  If $(\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.)$ in $\lambda_{\underline{a}.}\rho_{\underline{r}.}G$, then $P(b.|\rho\underline{a}. = a., \rho\underline{r}. = r., s.) = P(b.|a., \rho\underline{r}. = r., s.)$.

- **Rule 3:** Insertion and deletion of actions $(\rho\underline{a}. = a. \leftrightarrow 1)$

  If $(\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.)$ in $\rho_{\underline{a}.-an(\underline{s}.)}\rho_{\underline{r}.}G$, then $P(b.|\rho\underline{a}. = a., \rho\underline{r}. = r., s.) = P(b.|\rho\underline{r}. = r., s.)$.

These rules have been proven to be sufficient for removing all do operators from an expression for which it is possible to do so.

Next we discuss two theorems that can be proven using do-calculus: the backdoor and the front-door adjustment theorems.

We say that we are **adjusting or controlling a variable** $\underline{a}$ if we condition a probability on $\underline{a}$ and then we average that probability over $\underline{a}$. The backdoor theorem does one adjustment and the front-door theorem does two.

Backdoor Adjustment

See Chapter 1 for examples of the use of the backdoor adjustment theorem. In this section, we shall mainly be concerned with proving this theorem using do-calculus.

Suppose we have access to data that allows us to estimate a probability distribution $P(x., y., z.)$. Hence, the variables $\underline{x}., \underline{y}., \underline{z}.$ are all observed (i.e, not hidden). Then we say that the backdoor $\underline{z}.$ satisfies the **backdoor criterion** relative to $(\underline{x}., \underline{y}.)$ if

1. All paths from $\underline{x}.$ to $\underline{y}.$ that start with an arrow pointing into $\underline{x}.$, are blocked by $\underline{z}.$.

2. $\underline{z}. \notin de(\underline{x}.)$.

Claim 3 *Backdoor Adjustment*

If $\underline{z}.$ satisfies the backdoor criterion relative to $(\underline{x}., \underline{y}.)$, then

$$P(y.|\rho\underline{x}. = x.) = \sum_{z.} P(y.|x., z.)P(z.) \tag{9.8}$$

$$= \sum_{z.} \left\{ \begin{tikzcd} \underline{z}. = z. \arrow[dr] & \\ \underline{x}. = x. \arrow[r] & \underline{y}. \end{tikzcd} \right\} \tag{9.9}$$

proof:

For simplicity, let us omit the dots from the multinodes. If z satisfies the backdoor criterion relative to $(\underline{x}, \underline{y})$, then $\underline{x}, \underline{y}, \underline{z}$ must have the following structure.

$$\begin{tikzcd} \underline{z} \arrow[d] \arrow[dr] & \\ \underline{x} \arrow[r] & \underline{y} \end{tikzcd} \tag{9.10}$$

$P(y|\rho\underline{x} = x) =$

$= \sum_m P(y|\rho\underline{x} = x, z)P(z|\rho\underline{x} = x)$

by Probability Axioms

$= \sum_P (y|x, z)P(z|\rho\underline{x} = x)$

$P(y|\rho\underline{x} = x, z) \to P(y|x, z)$

by Rule 2: If $(\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.)$ in $\lambda_{\underline{a}.}\rho_{\underline{r}.}G$, then $P(b.|\rho\underline{a}. = a., \rho\underline{r}. = r., s.) = P(b.|a., \rho\underline{r}. = r., s.)$.

$\underline{y} \perp \underline{x}|\underline{z}$ in $\lambda_{\underline{x}}G$
$$\begin{tikzcd} \underline{z} \arrow[d] \arrow[dr] & \\ \underline{x} & \underline{y} \end{tikzcd}$$

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

$P(z|\rho\underline{x} = x) \to P(z)$

by Rule 3: If $(\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.)$ in $\rho_{\underline{a}.-an(\underline{s}.)}\rho_{\underline{r}.}G$, then $P(b.|\rho\underline{a}. = a., \rho\underline{r}. = r., s.) = P(b.|\rho\underline{r}. = r., s.)$.

$\underline{z} \perp \underline{x}$ in $\rho_{\underline{x}}G$
$$\begin{tikzcd} \underline{z} \arrow[dr] & \\ \underline{x} \arrow[r] & \underline{y} \end{tikzcd}$$

(9.11)

QED

Note that the backdoor adjustment formula can be written as

$$
\begin{aligned}
P(y.|\rho\underline{x}. = x.) &= \sum_{z.} P(y.|x., z.)P(z.) \qquad (9.12) \\
&= \sum_{z.} \frac{P(y., x., z.)}{P(x.|z.)} \qquad (9.13)
\end{aligned}
$$

$P(x.|z.)$ is called the **propensity score**, and one can approximate it to get an approximation to $P(y|\rho\underline{x} = x)$.

Front Door Adjustment

See Chapter 13 for examples of the use of the front-door adjustment theorem. In this section, we shall mainly be concerned with proving this theorem using do-calculus.

Suppose we have access to data that allows us to estimate a probability distribution $P(x., m., y.)$. Hence, the variables $\underline{x}., \underline{m}., \underline{y}.$ are all observed (i.e, not hidden). Then we say that the front-door $\underline{m}.$ satisfies the **front-door criterion** relative to $(\underline{x}., \underline{y}.)$ if

1. All directed paths from $\underline{x}.$ to $\underline{y}.$ are intercepted by (i.e., have a node in) $\underline{m}.$.
2. All backdoor paths from $\underline{x}.$ to $\underline{m}.$ are blocked.
3. All backdoor paths from on $\underline{m}.$ to $\underline{y}.$ are blocked by $\underline{x}.$.

Claim 4 *Front-Door Adjustment*

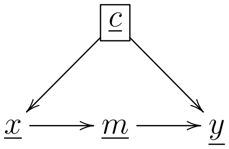
If $\underline{m}.$ satisfies the front-door criterion relative to $(\underline{x}., \underline{y}.)$, and $P(x., m.) > 0$, then

$$
P(y.|\rho\underline{x}. = x.) = \sum_{m.} \underbrace{\left[\sum_{x.'} P(y.|x'., m.)P(x'.)\right]}_{P(y.|\rho\underline{m}.=m.)} \underbrace{P(m.|x.)}_{P(m.|\rho\underline{x}.=x.)} \qquad (9.14)
$$

$$
= \sum_{m.,x.'} \left\{
\begin{tikzcd}
 & \underline{x}. = x.' \arrow[dr] & \\
\underline{x}. = x. \arrow[r] & \underline{m}. = m. \arrow[r] & \underline{y}.
\end{tikzcd}
\right\} \qquad (9.15)
$$

proof:

For simplicity, let us omit the dots from the multinodes. If $\underline{m}$ satisfies the front-door criterion relative to $(\underline{x}, \underline{y})$, then $\underline{x}, \underline{m}, \underline{y}$ must have the following structure, where node $\underline{c}$ is hidden.



(9.16)

Continues in next page.

$P(y|\rho\underline{x} = x) =$

$= \sum_m P(y|\rho\underline{x} = x, m)P(m|\rho\underline{x} = x)$

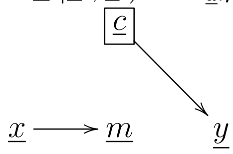
by Probability Axioms

$= \sum_m P(y|\rho\underline{x} = x, \rho\underline{m} = m)P(m|\rho\underline{x} = x)$

$P(y|\rho\underline{x} = x, m) \to P(y|\rho\underline{x} = x, \rho m = m)$

by Rule 2: If ($\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.$) in $\lambda_{\underline{a}.}\rho_{\underline{r}.}G$, then $P(b.|\rho\underline{a}. = a., \rho\underline{r}. = r., s.) = P(b.|a., \rho\underline{r}. = r., s.)$.

$\underline{y} \perp \underline{m}|\underline{x}$ in $\lambda_{\underline{m}}\rho_{\underline{x}}G$

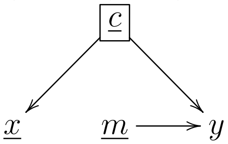


$= \sum_m P(y|\rho\underline{x} = x, \rho\underline{m} = m)P(m|x)$

$P(m|\rho\underline{x} = x) \to P(m|x)$

by Rule 2: If ($\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.$) in $\lambda_{\underline{a}.}\rho_{\underline{r}.}G$, then $P(b.|\rho\underline{a}. = a., \rho\underline{r}. = r., s.) = P(b.|a., \rho\underline{r}. = r., s.)$.

$\underline{m} \perp \underline{x}$ in $\lambda_{\underline{x}}G$

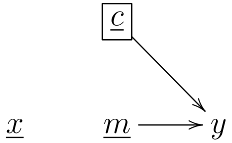


$= \sum_m P(y|\rho\underline{m} = m)P(m|x)$

$P(y|\rho\underline{x} = x, \rho\underline{m} = m) \to P(y|\rho\underline{m} = m)$

by Rule 3: If ($\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.$) in $\rho_{\underline{a}.-an(\underline{s}.)}\rho_{\underline{r}.}G$, then $P(b.|\rho\underline{a}. = a., \rho\underline{r}. = r., s.) = P(b.|\rho\underline{r}. = r., s.)$.

$\underline{y} \perp \underline{x}|\underline{m}$ in $\rho_{\underline{x}}\rho_{\underline{m}}G$



$= \sum_{x'} \sum_m P(y|\rho\underline{m} = m, x')P(x'|\rho\underline{m} = m)P(m|x)$

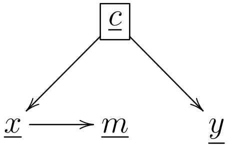
by Probability Axioms

$= \sum_{x'} \sum_m P(y|m, x')P(x'|\rho\underline{m} = m)P(m|x)$

$P(y|\rho\underline{m} = m, x') \to P(y|m, x')$

by Rule 2: If ($\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.$) in $\lambda_{\underline{a}.}\rho_{\underline{r}.}G$, then $P(b.|\rho\underline{a}. = a., \rho\underline{r}. = r., s.) = P(b.|a., \rho\underline{r}. = r., s.)$.

$\underline{y} \perp \underline{m}|\underline{x}$ in $\lambda_{\underline{m}}G$

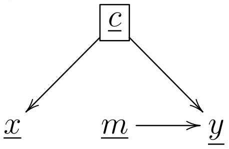


$= \sum_{x'} \sum_m P(y|m, x')P(x')P(m|x)$

$P(x'|\rho\underline{m} = m) \to P(x')$

by Rule 3: If ($\underline{b}. \perp \underline{a}.|\underline{r}., \underline{s}.$) in $\rho_{\underline{a}.-an(\underline{s}.)}\rho_{\underline{r}.}G$, then $P(b.|\rho\underline{a}. = a., \rho\underline{r}. = r., s.) = P(b.|\rho\underline{r}. = r., s.)$.

$\underline{x} \perp \underline{m}$ in $\rho_{\underline{m}}G$



(9.17)

QED

Chapter 13

Front-door Adjustment

The front-door (FD) adjustment theorem is proven in Chapter 9 from the rules of do-calculus. The goal of this chapter is to give examples of the use of that theorem. We will restate the theorem in this chapter, sans proof. There is no need to understand the theorem's proof in order to use it. However, you will need to skim Chapter 9 in order to familiarize yourself with the notation used to state the theorem. This chapter also assumes that you are comfortable with the rules for checking for d-separation. Those rules are covered in Chapter 10.

Suppose we have access to data that allows us to estimate a probability distribution $P(x., m., y.)$. Hence, the variables $\underline{x}., \underline{m}., \underline{y}.$ are all observed (i.e, not hidden). Then we say that the front-door $\underline{m}.$ satisfies the **front-door criterion** relative to $(\underline{x}., \underline{y}.)$ if

1. All directed paths from $\underline{x}.$ to $\underline{y}.$ are intercepted by (i.e., have a node in) $\underline{m}.$.

2. All backdoor paths from $\underline{x}.$ to $\underline{m}.$ are blocked.

3. All backdoor paths from on $\underline{m}.$ to $\underline{y}.$ are blocked by $\underline{x}.$.

Claim 5 *Front-Door Adjustment*
If $\underline{m}.$ satisfies the front-door criterion relative to $(\underline{x}., \underline{y}.)$, and $P(x., m.) > 0$, then

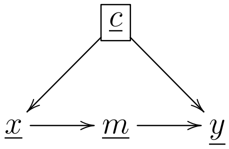
$$P(y.|\rho\underline{x}. = x.) = \sum_{m.} \underbrace{\left[\sum_{x.'} P(y.|x'., m.)P(x'.)\right]}_{P(y.|\rho\underline{m}.=m.)} \underbrace{P(m.|x.)}_{P(m.|\rho\underline{x}.=x.)} \tag{13.1}$$

$$= \sum_{m., x.'} \left\{ \begin{tikzcd} & \underline{x}. = x.' \arrow[dr] & \\ \underline{x}. = x. \arrow[r] & \underline{m}. = m. \arrow[r] & \underline{y}. \end{tikzcd} \right\} \tag{13.2}$$

proof: See Chapter 9
QED

Examples

1.



(13.3)

If $\underline{x}. = \underline{x}, \underline{m}. = \underline{m}$ and $\underline{y}. = \underline{y}$, then the FD criterion is satisfied.

Chapter 21

Linear Deterministic Bnets with Exogenous Noise

In this chapter, we will consider bnets which were referred to, prior to the invention of bnets, as: Sewall Wright's Path Analysis (PA) and linear Structural Equations Model (SEM). Judea Pearl in his books calls them linear Structural Causal Models (SCM), because they are very convenient for doing causal analysis. We will follow Judea's convention and refer to them as scum.

To build a SCM, start with a deterministic bnet G. Now add to each node $\underline{a}$ of G a root node $\underline{U}_{\underline{a}}$ pointing into $\underline{a}$ only. The nodes $\underline{U}_{\underline{a}}$ are called the **exogenous (external) variables**. The exogenous variables make their children noisy. Their TPMs are priors and are assumed to be unobserved. Since they are root nodes, they are mutually independent. When drawing bnets for SCM, we will never draw the exogenous nodes, leaving them implicit.

Examples:

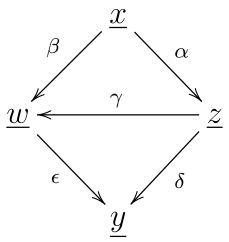


Figure 21.1: Example of a linear SCM wherein $\underline{x}$ splits into two nodes $\underline{z}$ and $\underline{w}$, then merges into node $\underline{y}$. There is also an arrow $\underline{z} \rightarrow \underline{w}$. Exogenous nodes are not shown.

1. The TPMs, printed in blue, for the nodes of the bnet Fig.21.1, are as follows.

$$P(y|w, z, U_{\underline{y}}) = \mathbb{1}(y = \epsilon w + \delta z + U_{\underline{y}}) \tag{21.1}$$

$$P(w|x, z, U_{\underline{w}}) = \mathbb{1}(w = \beta x + \gamma z + U_{\underline{w}}) \tag{21.2}$$

$$P(z|x, U_{\underline{z}}) = \mathbb{1}(z = \alpha x + U_{\underline{z}}) \tag{21.3}$$

$$P(x|U_{\underline{x}}) = \mathbb{1}(x = U_{\underline{x}}) \tag{21.4}$$

Hence,

$$\begin{aligned}
y &= \epsilon w + \delta z + U_{\underline{y}} && (21.5)\\
&= \epsilon(\beta x + \gamma z + U_{\underline{w}}) + \delta z + U_{\underline{y}} && (21.6)\\
&= (\epsilon\gamma + \delta)z + \epsilon\beta x + \epsilon U_{\underline{w}} U_{\underline{y}} && (21.7)\\
&= (\epsilon\gamma + \delta)z + \epsilon\beta U_{\underline{x}} + \epsilon U_{\underline{w}} U_{\underline{y}} && (21.8)
\end{aligned}$$

and

$$\left(\frac{\partial y}{\partial z}\right)_{U_. - U_{\underline{z}}} = \epsilon\gamma + \delta \ , \tag{21.9}$$

where the partial derivative holds fixed all exogenous variables except $U_{\underline{z}}$. Note that this partial derivative is a sum of terms, and that each of those terms represents a different causal path (one with all arrows pointing in the same direction, a Markov chain) from $\underline{z}$ to $\underline{y}(\underline{z})$. This is a general property of linear SCM diagrams.

2. Our next example uses the notation $\langle \underline{x}, \underline{y} \rangle$ for covariances of any two random variables $\underline{x}, \underline{y}$. The $\langle \underline{x}, \underline{y} \rangle$ notation is defined in the Notational Conventions Chapter 0.3.

   Consider a bnet with deterministic nodes $\underline{x}_. = (\underline{x}_k)_{k=0,1,\ldots nx-1}$ and corresponding exogenous nodes $\underline{U}_. = (\underline{U}_k)_{k=0,1,\ldots nx-1}$. Assume $\langle \underline{U}_i, \underline{U}_j \rangle = 0$ if $i \neq j$. The **structural coefficient** $\alpha_{j|i} > 0$ measures the strength of the connection $\underline{x}_i \rightarrow \underline{x}_j$.

   - **Fully connected graph with** $nx = 2$

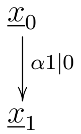


Figure 21.2: Fully connected graph with two $\underline{x}_j$ nodes (exogenous nodes $\underline{U}_j$ not shown).

Note that

$$\begin{aligned}
\underline{x}_0 &= \underline{U}_0 && (21.10a)\\
\underline{x}_1 &= \alpha_{1|0}\underline{x}_0 + \underline{U}_1 && (21.10b)
\end{aligned}$$

. Eqs.21.10 constitute a system of 2 linear equations in 2 unknowns (the $\underline{x}$'s) so we can solve for the $\underline{x}$'s in terms of the α's and $\underline{U}$'s.

Note also that

$$\langle \underline{x}_1, \underline{x}_0 \rangle = \alpha_{1|0} \langle \underline{x}_0, \underline{x}_0 \rangle \; . \tag{21.11}$$

Thus, $\alpha_{1|0}$ can be estimated from the covariances $\langle \underline{x}_i, \underline{x}_j \rangle$.

- **Fully connected graph with** $nx = 3$

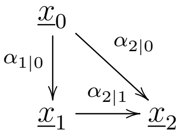


Figure 21.3: Fully connected graph with three $\underline{x}_j$ nodes (exogenous nodes $\underline{U}_j$ not shown).

Note that

$$\begin{aligned}
\underline{x}_0 &= \underline{U}_0 && (21.12a)\\
\underline{x}_1 &= \alpha_{1|0}\underline{x}_0 + \underline{U}_1 && (21.12b)\\
\underline{x}_2 &= \alpha_{2|1}\underline{x}_1 + \alpha_{2|0}\underline{x}_0 + \underline{U}_2 \; . && (21.12c)
\end{aligned}$$

Eqs.21.12 constitute a system of 3 linear equations in 3 unknowns (the $\underline{x}$'s) so we can solve for the $\underline{x}$'s in terms of the α's and $\underline{U}$'s.

Note also that

$$\begin{aligned}
\langle \underline{x}_1, \underline{x}_0 \rangle &= \alpha_{1|0} \langle \underline{x}_0, \underline{x}_0 \rangle && (21.13a)\\
\langle \underline{x}_2, \underline{x}_0 \rangle &= \alpha_{2|1} \langle \underline{x}_1, \underline{x}_0 \rangle + \alpha_{2|0} \langle \underline{x}_0, \underline{x}_0 \rangle && (21.13b)\\
\langle \underline{x}_2, \underline{x}_1 \rangle &= \alpha_{2|1} \langle \underline{x}_1, \underline{x}_1 \rangle + \alpha_{2|0} \langle \underline{x}_0, \underline{x}_1 \rangle && (21.13c)
\end{aligned}$$

Eqs.21.13 constitute a system of 3 linear equations in 3 unknowns (the α's) so we can solve solve for the α's in terms of covariances $\langle \underline{x}_i, \underline{x}_j \rangle$. This gives an estimate for the α's.

Bibliography

[1] Judea Pearl. *Probabilistic Inference in Intelligent Systems.* Morgan Kaufmann, 1988.

[2] Judea Pearl. *Causality: Models, Reasoning, and Inference, Second Edition.* Cambridge University Press, 2013.

[3] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer.* John Wiley & Sons, 2016.

[4] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect.* Basic Books, 2018.

[5] Wikipedia. Boolean algebra. `https://en.wikipedia.org/wiki/Boolean_algebra`.

[6] Robert R. Tucci. Bell's inequalities for Bayesian statisticians. blog post in blog Quantum Bayesian Networks, `https://qbnets.wordpress.com/2008/09/19/bells-inequaties-for-bayesian-statistician/`.

[7] Wikipedia. Binary decision diagram. `https://en.wikipedia.org/wiki/Binary_decision_diagram`.

[8] Wikipedia. Expectation maximization. `https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm`.

[9] Wikipedia. k-means clustering. `https://en.wikipedia.org/wiki/K-means_clustering`.

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, David Warde-Farley Bing Xu, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. `https://arxiv.org/abs/1406.2661`.

[11] Wikipedia. Hidden Markov model. `https://en.wikipedia.org/wiki/Hidden_Markov_model`.

[12] Gregory Nuel. Tutorial on exact belief propagation in Bayesian networks: from messages to algorithms. `https://arxiv.org/abs/1201.4724`.

[13] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988. `http://www.eecis.udel.edu/~shatkay/Course/papers/Lauritzen1988.pdf`.

[14] Wikipedia. Junction tree algorithm. https://en.wikipedia.org/wiki/Junction_tree_algorithm.

[15] Cecil Huang and Adnan Darwiche. Inference in belief networks: A procedural guide. *International journal of approximate reasoning*, 15(3):225–263, 1996. http://www.ar-tiste.com/Huang-Darwiche1996.pdf.

[16] Robert R. Tucci. Quantum Fog. https://github.com/artiste-qb-net/quantum-fog.

[17] Wikipedia. Kalman filter. https://en.wikipedia.org/wiki/Kalman_filter.

[18] Wikipedia. Markov blanket. https://en.wikipedia.org/wiki/Markov_blanket.

[19] Wikipedia. Monte Carlo methods. https://en.wikipedia.org/wiki/Category:Monte_Carlo_methods.

[20] Wikipedia. Inverse transform sampling. https://en.wikipedia.org/wiki/Inverse_transform_sampling.

[21] Wikipedia. Rejection sampling. https://en.wikipedia.org/wiki/Rejection_sampling.

[22] Dan Bendel. Metropolis-Hastings: A comprehensive overview and proof. https://similarweb.engineering/mcmc/.

[23] Wikipedia. Metropolis-Hastings method. https://en.wikipedia.org/wiki/Metropolis%E2%80%93Hastings_algorithm.

[24] Wikipedia. Gibbs sampling. https://en.wikipedia.org/wiki/Gibbs_sampling.

[25] Wikipedia. Importance sampling. https://en.wikipedia.org/wiki/Importance_sampling.

[26] Judea Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. https://www.aaai.org/Papers/AAAI/1982/AAAI82-032.pdf, 1982.

[27] Wikipedia. Belief propagation. https://en.wikipedia.org/wiki/Belief_propagation.

[28] Richard E Neapolitan. *Learning bayesian networks.* Pearson Prentice Hall, 2004.

[29] Nitish Srivastava, G E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf.

[30] Wikipedia. Non-negative matrix factorization. https://en.wikipedia.org/wiki/Non-negative_matrix_factorization.

[31] theinvestorsbook.com. Pert analysis. https://theinvestorsbook.com/pert-analysis.html.

[32] Wikipedia. Program evaluation and review technique. https://en.wikipedia.org/wiki/Program_evaluation_and_review_technique.

[33] Andrew Ng. Lecture at deeplearning.ai on recurrent neural networks. http://www.ar-tiste.com/ng-lec-rnn.pdf.

[34] Wikipedia. Long short term memory. https://en.wikipedia.org/wiki/Long_short-term_memory.

[35] Wikipedia. Gated recurrent unit. https://en.wikipedia.org/wiki/Gated_recurrent_unit.

[36] Charles Fox, Neil Girdhar, and Kevin Gurney. A causal bayesian network view of reinforcement learning. https://www.aaai.org/Papers/FLAIRS/2008/FLAIRS08-030.pdf".

[37] Sergey Levine. Course CS 285 at UC Berkeley, Deep reinforcement learning. http://rail.eecs.berkeley.edu/deeprlcourse/.

[38] ReliaSoft. System analysis reference. http://reliawiki.org/index.php/System_Analysis_Reference.

[39] W. E. Vesely, F. F. Goldberg, N. H. Roberts, and D. F. Haasl. Fault tree handbook nureg-0492. https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0492/.

[40] Wikipedia. Simpson's paradox. https://en.wikipedia.org/wiki/Simpson's_paradox.

[41] Robert J. McEliece, David J. C. MacKay, and Jung-Fu Cheng. Turbo decoding as an instance of Pearls belief propagation algorithm. http://authors.library.caltech.edu/6938/1/MCEieeejstc98.pdf.

[42] Wikipedia. Variational bayesian methods. https://en.wikipedia.org/wiki/Variational_Bayesian_methods.