Bayesuvius,
a small visual dictionary of Bayesian Networks

Robert R. Tucci
www.ar-tiste.xyz

October 5, 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 **directed graph** $G = (V, E)$ consists of two sets, V and E. V contains the **vertices (nodes)** and E contains the **edges (arrows)**. An arrow $a \to b$ is an ordered pair (a, b) where $a, b \in V$. The **parents** of a node x are those nodes a such that there are arrows $a \to x$. The **children** of a node x are those nodes b such that there are arrows $x \to b$. The **neighbors** of a node x is the set of parents and children of x. A **path** is a set of nodes that are connected by arrows, so that all nodes have 1 or 2 neighbors. A **directed path** is a path in which all the arrows point in the same direction. A **loop** is a closed path;i.e., a path in which all nodes have exactly 2 neighbors. A **cycle** is a directed loop. A **Directed Acyclic Graph (DAG)** is a directed graph that has no cycles.

A **fully connected directed graph** is a directed graph in which every node has all other nodes as neighbors. Figs.3 and 4 show 2 different ways of drawing the same directed graph, a fully connected graph with 4 nodes. Note that a convenient way to label the nodes of a fully connected directed graph with N nodes is to point arrows from $\underline{x}_j$ to $\underline{x}_k$ where $j = 0, 1, \ldots, N-1$ and $k = j+1, j+2, \ldots, N-1$.

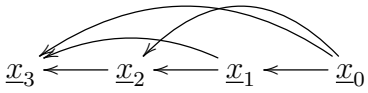


Figure 3: Fully connected directed graph with 4 nodes, drawn as a line.

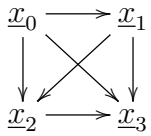


Figure 4: Fully connected directed graph with 4 nodes, drawn as a square.

A **Bayesian network (bnet)** consists of a DAG 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)** $\underline{a}. = (\underline{a}_j)_{j=0,1,...,na-1}$. 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

$$P(x.) = \prod_j P(\underline{x}_j = x_j|(\underline{x}_k = x_k)_{k:\underline{x}_k \in pa(\underline{x}_j)}) \ . \tag{1}$$

Note that for a fully connected bnet with N nodes, Eq.(1) becomes

$$P(x.) = \prod_{j=0}^{N-1} P(x_j|(x_k)_{k=j-1,j-2,...,0}) \ . \tag{2}$$

For example, if $N = 4$, Eq.(2) becomes

$$P(x_0, x_1, x_2, x_3) = P(x_3|x_2, x_1, x_0)P(x_2|x_1, x_0)P(x_1|x_0)P(x_0) \ . \tag{3}$$

We see that Eq.(2) is just the chain rule for conditional probabilities.

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,

$$\sum_{x \in S_{\underline{x}}} P_{\underline{x}}(x) = 1 \tag{4}$$

$$P_{\underline{x},\underline{y}}(x, y) = P(\underline{x} = x, \underline{y} = y) = P(x, y) \tag{5}$$

$$P_{\underline{x}|\underline{y}}(x|y) = P(\underline{x} = x|\underline{y} = y) = P(x|y) = \frac{P(x, y)}{P(y)} \tag{6}$$

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

$$\delta(x, y) = \begin{cases} 1 \text{ if } x = y \\ 0 \text{ if } x \neq y \end{cases} \tag{7}$$

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

$$\int_{-\infty}^{+\infty} dx\ \delta(x - y) f(x) = f(y) \tag{8}$$

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

$$\frac{1}{(b-a)} \int_{[a,b]} dx\ f(x) \to \frac{\Delta x}{(b-a)} \sum_{x \in S_{\underline{x}}} f(x) = \frac{1}{n_{\underline{x}}} \sum_{x \in S_{\underline{x}}} f(x)\ . \tag{9}$$

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

$$\int_{[a,b]} dx\ \delta(x-y) f(x) = f(y) \to \sum_{x \in S_{\underline{x}}} \delta(x,y) f(x) = f(y)\ . \tag{10}$$

Indicator function (aka Truth function):

$$\mathbb{1}(\mathcal{S}) = \begin{cases} 1 \text{ if } \mathcal{S} \text{ is true} \\ 0 \text{ if } \mathcal{S} \text{ is false} \end{cases} \tag{11}$$

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

$$\vec{x} = (x[0], x[1], x[2] \ldots, x[nsam(\vec{x}) - 1]) = x[:] \tag{12}$$

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

$$x[i] \sim P_{\underline{x}} \ \text{ (i.e. } P_{\underline{x[i]}} = P_{\underline{x}}) \tag{13}$$

$$P(\underline{x} = x) = \frac{1}{nsam(\vec{x})} \sum_i \mathbb{1}(x[i] = x) \tag{14}$$

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

$$\sum_x P(\underline{x} = x) f(x) = \frac{1}{nsam(\vec{x})} \sum_i f(x[i]) \tag{15}$$

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.

$$P(\vec{x}) = \prod_i P(x[i]) \tag{16}$$

$$\sum_{\vec{x}} = \prod_i \sum_{x[i]} \tag{17}$$

$$\partial_{\vec{x}} = [\partial_{x[0]}, \partial_{x[1]}, \partial_{x[2]}, \ldots, \partial_{x[nsam(\vec{x})-1]}] \tag{18}$$

$$
\begin{aligned}
P(\vec{x}) &\approx [\prod_x P(x)^{P(x)}]^{nsam(\vec{x})} && (19)\\
&= e^{nsam(\vec{x})\sum_x P(x)\ln P(x)} && (20)\\
&= e^{-nsam(\vec{x})H(P_{\underline{x}})} && (21)
\end{aligned}
$$

$$f^{[1,\partial_x,\partial_y]}(x,y) = [f, \partial_x f, \partial_y f] \tag{22}$$

$$f^{+} = f^{[1,\partial_x,\partial_y]} \tag{23}$$

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

- Entropy:

$$H(p) = -\sum_x p(x)\ln p(x) \geq 0 \tag{24}$$

- Kullback-Liebler divergence:

$$D_{KL}(p \parallel q) = \sum_x p(x)\ln\frac{p(x)}{q(x)} \geq 0 \tag{25}$$

- Cross entropy:

$$
\begin{aligned}
CE(p \to q) &= -\sum_x p(x)\ln q(x) && (26)\\
&= H(p) + D_{KL}(p \parallel q) && (27)
\end{aligned}
$$

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

$$\mathcal{N}(\mu,\sigma^2)(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{28}$$

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

$$\mathcal{U}(a,b)(x) = \frac{1}{b-a} \tag{29}$$

Expected Value and Variance

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

$$E_{\underline{x}}[f(\underline{x})] = E_{x\sim P(x)}[f(x)] = \sum_x P(x)f(x) \tag{30}$$

$$
\begin{aligned}
Var_{\underline{x}}[f(\underline{x})] &= E_{\underline{x}}\left[(f(\underline{x}) - E_{\underline{x}}[f(\underline{x})])^2\right] && (31)\\
&= E_{\underline{x}}[f(\underline{x})^2] - (E_{\underline{x}}[f(\underline{x})])^2 && (32)
\end{aligned}
$$

$$E[\underline{x}] = E_{\underline{x}}[\underline{x}] \tag{33}$$

$$Var[\underline{x}] = Var_{\underline{x}}[\underline{x}] \tag{34}$$

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

$$E_{\underline{x}|\underline{y}}[f(\underline{x}, \underline{y})] = \sum_x P(x|\underline{y}) f(x, \underline{y}) \ , \tag{35}$$

$$E_{\underline{x}|\underline{y}=y}[f(\underline{x}, y)] = E_{\underline{x}|y}[f(\underline{x}, y)] = \sum_x P(x|y) f(x, y) \ . \tag{36}$$

Note that

$$\begin{aligned} E_{\underline{y}}[E_{\underline{x}|\underline{y}}[f(\underline{x}, \underline{y})]] &= \sum_{x,y} P(x|y) P(y) f(x, y) && (37) \\ &= \sum_{x,y} P(x, y) f(x, y) && (38) \\ &= E_{\underline{x},\underline{y}}[f(\underline{x}, \underline{y})] \ . && (39) \end{aligned}$$

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*

$$Var_{\underline{x},\underline{y}}(f) = E_{\underline{y}}[Var_{\underline{x}|\underline{y}}(f)] + Var_{\underline{y}}(E_{\underline{x}|\underline{y}}[f]) \ . \tag{40}$$

In particular,

$$Var_{\underline{x}}(x) = E_{\underline{y}}[Var_{\underline{x}|\underline{y}}(x)] + Var_{\underline{y}}(E_{\underline{x}|\underline{y}}[x]) \ . \tag{41}$$

proof:

Let

$$A = \sum_y P(y) \left(\sum_x P(x|y) f \right)^2 \ . \tag{42}$$

Then

$$\begin{aligned} Var_{\underline{x},\underline{y}}(f) &= \sum_{x,y} P(x, y) f^2 - \left(\sum_{x,y} P(x, y) f \right)^2 && (43) \\ &= \begin{cases} \sum_{x,y} P(x, y) f^2 - A \\ + \left(A - \left(\sum_{x,y} P(x, y) f \right)^2 \right) \end{cases} && (44) \end{aligned}$$

$$E_{\underline{y}}[Var_{\underline{x}|\underline{y}}(f)] = \sum_y P(y) \left(\sum_x P(x|y) f^2 - \left(\sum_x P(x|y) f \right)^2 \right) \tag{45}$$

$$= \sum_{x,y} P(x,y) f^2 - A \tag{46}$$

$$Var_{\underline{y}}(E_{\underline{x}|\underline{y}}[f]) = \sum_y P(y) \left(\sum_x P(x|y) f \right)^2 - \left(\sum_y P(y) \sum_x P(x|y) f \right)^2 \tag{47}$$

$$= A - \left(\sum_{x,y} P(x,y) f \right)^2 \tag{48}$$

QED

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

Mean value of $\underline{x}$

$$\langle \underline{x} \rangle = E_{\underline{x}}[\underline{x}] \tag{49}$$

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

$$\Delta \underline{x} = \underline{x} - \langle \underline{x} \rangle \tag{50}$$

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

$$\langle \underline{x}, \underline{y} \rangle = \langle \Delta \underline{x} \Delta \underline{y} \rangle = Cov(\underline{x}, \underline{y}) \tag{51}$$

Variance of $\underline{x}$

$$Var(\underline{x}) = \langle \underline{x}, \underline{x} \rangle \tag{52}$$

Standard deviation or $\underline{x}$

$$\sigma_{\underline{x}} = \sqrt{\langle \underline{x}, \underline{x} \rangle} \tag{53}$$

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

$$\rho_{\underline{x},\underline{y}} = \frac{\langle \underline{x}, \underline{y} \rangle}{\sqrt{\langle \underline{x}, \underline{x} \rangle \langle \underline{y}, \underline{y} \rangle}} \tag{54}$$

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

$$\text{sig}(x) = \frac{1}{1 + e^{-x}} \tag{55}$$

$\mathcal{N}(!a)$ 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.[1] for more info about this topic.

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

$$x \text{ or } y = x \vee y = x + y - xy \ , \tag{56}$$

$$x \text{ and } y = x \wedge y = xy \ , \tag{57}$$

and

$$\text{not } x = \overline{x} = 1 - x \ , \tag{58}$$

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

$$P(\overline{x}) = 1 - P(x) \ . \tag{59}$$

[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 $(x, y) \in \{0, 1\}^2$,

$$P(x \vee y) = P(x) + P(y) - P(x \wedge y) \ . \tag{60}$$

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.[2] 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[3].

2. In 2016, he released together with Glymour and Jewell, a less advanced "primer" on causality, Ref.[4].

3. In 2018, he released together with Mackenzie his lovely "The Book of Why", Ref.[5].

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 10 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 10 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 11.

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 adjustment 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. $\underline{z}. \cap de(\underline{x}.) = \emptyset$.

Claim 2 Backdoor Adjustment Theorem

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

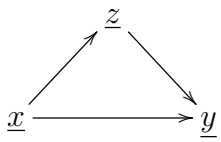
$$P(y.|\rho\underline{x}. = x.) = \sum_{z.} P(y.|x., z.)P(z.) \tag{1.1}$$

$$= \sum_{z.} \left\{ \begin{array}{ccc} \underline{z}. = z. & & \\ & \searrow & \\ \underline{x}. = x. & \longrightarrow & \underline{y}. \end{array} \right\} \tag{1.2}$$

proof: See Chapter 10
QED

Examples:

1.

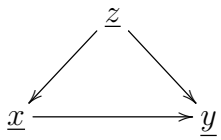


(1.3)

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.

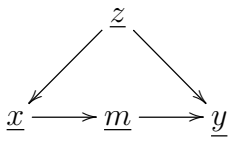


(1.5)

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 $\underline{x}$..

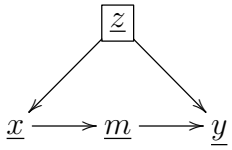
3.



(1.6)

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

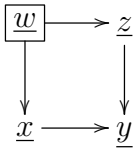
4.



(1.7)

BD criterion is impossible to satisfy if $\underline{x}. = \underline{x}, \underline{y}. = \underline{y}$. However, the front-door criterion can be satisfied. See Chapter 14.

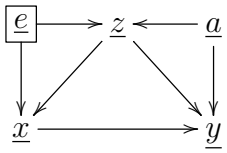
5.



(1.8)

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.

6.

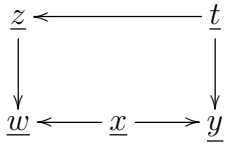


(1.9)

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.

7.



(1.10)

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

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

However, if for some reason we want to control $\underline{t}$, we can do so. We can't control $\underline{w}$ though, because $\underline{w} \in de(\underline{x})$. Thus, the BD criterion is satisfied if $\underline{x}. = \underline{x}$, $\underline{y}. = \underline{y}$ and $\underline{z}. = \underline{t}$. Therefore,

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

---

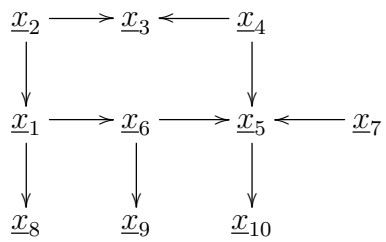
8. Discuss reasons why multiple possible sets $\underline{z}.$ that satisfy the BD criterion can be advantageous.

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

---

9. (Taken from online course notes Ref.[6])

Consider the bnet



(1.13)

If $\underline{x}_\cdot = \underline{x}_1$ and $\underline{y}_\cdot = \underline{x}_5$, find all possible adjustment multinodes $\underline{z}_\cdot$ that satisfy the BD criterion. Ans:

- $\emptyset$
- $\underline{x}_2$
- $\underline{x}_4$
- $\underline{x}_2, \underline{x}_4$
- $\underline{x}_2, \underline{x}_3$
- $\underline{x}_3, \underline{x}_4$
- $\underline{x}_2, \underline{x}_3, \underline{x}_4$

Add $\underline{x}_7$ to each of the previous 7 possible $\underline{z}_\cdot$. This gives a total of 14 possible adjustment multinodes $\underline{z}_\cdot$.

Chapter 7

Counterfactual Reasoning: COMING SOON

This chapter assumes that the reader has read Chapter 22 on linear deterministic systems with exogenous noise.

Let us repeat Eq.(22.21) from Chapter 22.

$$\underline{x} = A\underline{x} + \underline{u} \tag{7.1}$$

This equation represents the structural equations for a fully connected PA diagram.

Actually, we want to consider $nsam$ copies of equation Eq.(7.1)

$$\underline{x}[s] = A\underline{x}[s] + \underline{u}[s] \ , \tag{7.2}$$

where each $s = 0, 1, \ldots, nsam-1$. Each s represents a separate individual or "unit" in a population.

1. Solve Eqs.7.2 for $\underline{u}$.

$$\underline{u}(\underline{x}, A) = (1 - A)\underline{x} \tag{7.3}$$

2. Modify Eqs.7.2 by replacing equation for $\underline{x}_{i^*}$ by

$$\underline{x}_{i^*} = a \tag{7.4}$$

for some $a \in S_{\underline{x}_{i^*}}$. Modify the PA diagram correspondingly by deleting all arrows entering node $\underline{x}_{i^*}$. This is an intervention similar to those done in do-calculus.

Define A^* by

$$A^*_{r,c} = \begin{cases} A_{r,c} & \text{if row } r \neq i^* \\ 0 & \text{if row } r = i^* \end{cases} \tag{7.5}$$

Define $\underline{u}^*$ by

$$\underline{u}^*_r = \begin{cases} \underline{u}_r(\underline{x}, A) & \text{if row } r \neq i^* \\ a & \text{if row } r = i^* \end{cases} \tag{7.6}$$

Define new random variables $\underline{x}^*$ that satisfy

$$\underline{x}^* = A^* \underline{x}^* + \underline{u}^* \ . \tag{7.7}$$

Thus,

$$\underline{x}^* = (1 - A^*)^{-1} \underline{u}^* \tag{7.8}$$

$$P(x_i^* | x_{<i}^*) = \mathbb{1}(x_i^* = \sum_{k<i} \alpha_{i|k}^* x_k^* + u_i^*) \ . \tag{7.9}$$

$$P(x^*.) = \prod_i P(x_i^* | x_{<i}^*) \tag{7.10}$$

Chapter 10

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 inside a black box) or (2) making the arrows coming out of them dashed. Accordingly, the 3 diagrams in Fig.10.1 all mean the same thing. A **confounder node for** $\underline{x} \to \underline{y}$ (such as node $\underline{c}$ in Fig.10.1) is a hidden node with arrows pointing from it to both $\underline{x}$ and $\underline{y}$.

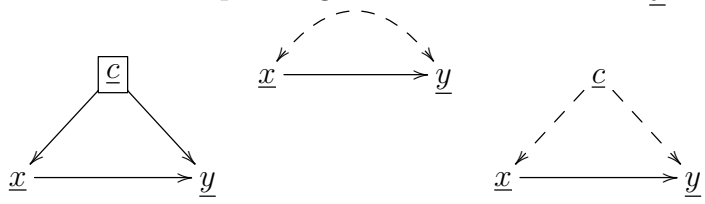


Figure 10.1: These 3 diagrams are equivalent. They mean that node $\underline{c}$ is hidden. Node $\underline{c}$ is implicit in the middle diagram.

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 converting $\underline{x}$ into a new node $\lambda\underline{x}$ that is a leaf node. $\rho_{\underline{x}}$ and $\lambda_{\underline{x}}$ are depicted in Fig.10.2.

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 $\underline{a}.$.

$$\rho_{\underline{a}.}G = \left[\prod_j \rho_{\underline{a}_j}\right] G \ , \quad \lambda_{\underline{a}.}G = \left[\prod_j \lambda_{\underline{a}_j}\right] G \ . \tag{10.1}$$

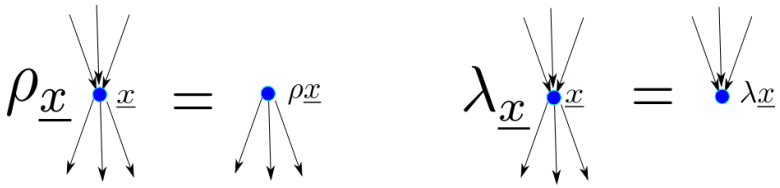


Figure 10.2: The operator $\rho_{\underline{x}}$ converts node $\underline{x}$ into a root node $\rho\underline{x}$. The operator $\lambda_{\underline{x}}$ converts node $\underline{x}$ into a leaf node $\lambda\underline{x}$.

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

$$P(X.) = \prod_j P(X_j|(X_k)_{k:\underline{X}_k \in pa(\underline{X}_j)}) \ . \tag{10.2}$$

Define an operator ρ that acts as follows[1]

$$\begin{aligned} P(X. - a.|\rho\underline{a}. = a.) &= \mathcal{N}(!(X. - a.)) \frac{P(X.)}{\prod_{j:\underline{X}_j \in \underline{a}.} P(X_j|(X_k)_{k:\underline{X}_k \in pa(\underline{X}_j)})} && (10.3) \\ &= \mathcal{N}(!(X. - a.)) \prod_{j:\underline{X}_j \notin \underline{a}.} P(X_j|(X_k)_{k:\underline{X}_k \in pa(\underline{X}_j)}) && (10.4) \\ &\neq P(X. - a.|\underline{a}. = a.) \ . && (10.5) \end{aligned}$$

For instance, for the bnet

$$\underline{x} \longrightarrow \underline{y} \tag{10.6}$$

with

$$P(x, y) = P(y|x)P(x) \ , \tag{10.7}$$

one has

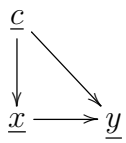
$$P(y|\rho\underline{x} = x) = P(y|x) \tag{10.8}$$

and

$$P(x|\rho\underline{y} = y) = P(x) \ . \tag{10.9}$$

This means that $\underline{x}$ causes $\underline{y}$ and $\underline{y}$ does not cause $\underline{x}$.

For the bnet



(10.10)

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

with

$$P(x, y, c) = P(y|x, c)P(x|c)P(c) \ , \tag{10.11}$$

one has

$$P(y, c|\rho\underline{x} = x) = P(y|x, c)P(c) \ . \tag{10.12}$$

Hence,

$$P(y|\rho\underline{x} = x) = \sum_c P(y|x, c)P(c) \ . \tag{10.13}$$

This is called **adjusting the parents of** $\underline{x}$.

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

$$P(b.|\rho\underline{a}. = a.) = \sum_{X.-a.-b.} P(X. - a.|\rho\underline{a}. = a.) \ , \tag{10.14}$$

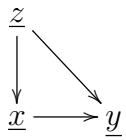
and for $\underline{s}. \subset \underline{X}. - \underline{a}. - \underline{b}.$, define

$$P(b.|\rho\underline{a}. = a., s.) = \frac{P(b., s.|\rho\underline{a}. = a.)}{P(s.|\rho\underline{a}. = a.)} \ . \tag{10.15}$$

$P(b.|\rho\underline{a}. = a., s.)$ 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**.

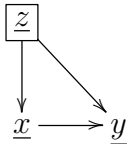
In $P(y|\rho\underline{x} = x)$, node $\underline{x}$ is turned into a root node. This guarantees that there is no confounding node connecting $\underline{x}$ and $\underline{y}$. Such confounding nodes are unwelcomed when calculating causal effects between the 2 variables $\underline{x}$ and $\underline{y}$ because they introduce non-causal correlations between the two. This is also what happens in a **Randomized Clinical Trial (RCT)**. In a RCT with treatment $\underline{x}$, the value of $\underline{x}$ for each patient is determined by a coin toss, effectively turning $\underline{x}$ into a root node. Hence, the do operator mimics a RCT.

$P(b.|\rho\underline{a}. = a., s.)$ 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 example, $P(y|\rho\underline{x} = x)$ is identifiable for the bnet



(10.16)

but it is non-identifiable for the bnet



(10.17)

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

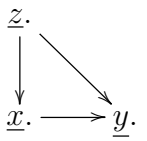
$$ACE = P(y = 1|\rho\underline{x} = 1) - P(y = 1|\rho\underline{x} = 0) \tag{10.18}$$

and the **Risk Difference (RD)** as

$$RD = P(y = 1|\underline{x} = 1) - P(y = 1|\underline{x} = 0) \ . \tag{10.19}$$

Parent Adjustment

Suppose that $\underline{x}., \underline{y}., \underline{z}.$ are disjoint multinodes and their union equals the totality of all nodes of a bnet. Suppose we have data available that allows us to estimate $P(x., y., z.)$. Hence, all nodes of the bnet are observable. Furthermore, suppose $\underline{z}. = pa(\underline{x}.)$. In other words, we are considering the bnet



. (10.20)

Then

$$P(y., z.|\rho\underline{x}. = x.) = P(y.|x., z.)P(z.) \tag{10.21}$$

so

$$P(y.|\rho\underline{x}. = x.) = \sum_{z.} P(y.|x., z.)P(z.) \ . \tag{10.22}$$

This is called **adjusting the parents** of $\underline{x}.$.

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 $\underline{a}$. More generally, we can adjust a whole multinode $\underline{a}.$ together.

Later on, we will introduce a generalization of this parent adjustment called the backdoor adjustment. In a backdoor adjustment, the adjusted multinode is not necessarily the parents of $\underline{x}.$, and $P(x., y., z.)$ need not represent the whole bnet.

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 11 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.

The backdoor theorem adjusts one multinode and the front-door theorem adjusts 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 adjustment 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. $\underline{z}. \cap de(\underline{x}.) = \emptyset$.

Motivation for BD criterion: Part 1 rules out paths from $\underline{x}$ to $\underline{y}$ containing a fork node (confounder) which, if not blocked by $\underline{z}.$, would introduce a non-causal correlation (confounder bias). Part 2 rules out a directed path from $\underline{x}$ to $\underline{y}$ that has a mediator node blocked by $\underline{z}.$ or a collider node unblocked by $\underline{z}.$.

Claim 3 *Backdoor Adjustment Theorem*
If $\underline{z}.$ satisfies the backdoor criterion relative to $(\underline{x}., \underline{y}.)$, then

$$P(y.|\rho\underline{x}. = x.) = \sum_{z.} P(y.|x., z.)P(z.) \tag{10.23}$$

$$= \sum_{z.} \left\{ \begin{array}{ccc} \underline{z}. = z. & & \\ & \searrow & \\ \underline{x}. = x. & \longrightarrow & \underline{y}. \end{array} \right\} \tag{10.24}$$

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.

$$\begin{array}{ccc} \underline{z} & \searrow & \\ \downarrow & & \\ \underline{x} & \longrightarrow & \underline{y} \end{array} \tag{10.25}$$

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

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

by Probability Axioms

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

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

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 $\lambda_{\underline{x}}G$

$$\begin{array}{ccc} \underline{z} & \searrow & \\ \downarrow & & \\ \underline{x} & & \underline{y} \end{array}$$

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

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

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 $\rho_{\underline{x}}G$

$$\begin{array}{ccc} \underline{z} & \searrow & \\ & & \\ \underline{x} & \longrightarrow & \underline{y} \end{array} \tag{10.26}$$

QED

Note that the backdoor adjustment formula can be written as

$$P(y.|\rho\underline{x}. = x.) = \sum_{z.} P(y.|x., z.)P(z.) \tag{10.27}$$

$$= \sum_{z.} \frac{P(y., x., z.)}{P(x.|z.)} \tag{10.28}$$

This assumes $P(x.|z.) \neq 0$ for all $x., z.$. This assumption is referred to as **positivity**, and is violated if $P(x.|z.) = \delta(x., x.(z.))$. $P(x.|z.)$ is called the **propensity score** of $x.$ given $z.$. This equation does **inverse probability weighting**. One can approximate $P(x.|z.)$ in this equation to get an approximation to $P(y|\rho\underline{x} = x)$.

Front Door Adjustment

See Chapter 14 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 adjustment 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 $\underline{x}.$.

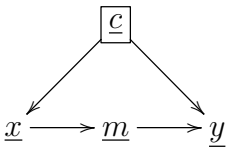
Claim 4 *Front-Door Adjustment Theorem*
If $\underline{m}.$ satisfies the front-door criterion relative to $(\underline{x}., \underline{y}.)$, and $P(x., m.) > 0$, then

$$P(y.|\rho\underline{x}. = x.) = \sum_{m.} \underbrace{\left[\sum_{x.'} P(y.|x'., m.)P(x'.)\right]}_{P(y.|\rho\underline{m}.=m.)} \underbrace{P(m.|x.)}_{P(m.|\rho\underline{x}.=x.)} \tag{10.29}$$

$$= \sum_{m.,x.'} \left\{ \begin{array}{ccc} & \underline{x}. = x.' & \\ & \searrow & \\ \underline{x}. = x. & \longrightarrow & \underline{m}. = m. \longrightarrow \underline{y}. \end{array} \right\} \tag{10.30}$$

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.



(10.31)

Continues in next page.

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

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

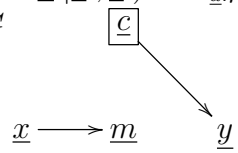
by Probability Axioms

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

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

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 $\lambda_{\underline{m}}\rho_{\underline{x}}G$

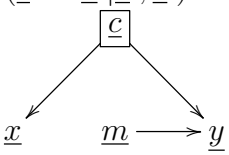


$= \sum_m P(y|\rho\underline{x} = x, \rho\underline{m} = m)P(m|x)$

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

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 $\lambda_{\underline{x}}G$

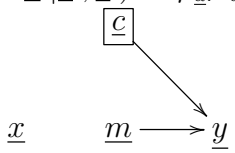


$= \sum_m P(y|\rho\underline{m} = m)P(m|x)$

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

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



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

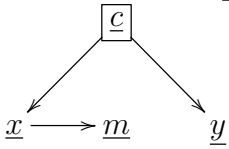
by Probability Axioms

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

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

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 $\lambda_{\underline{m}}G$

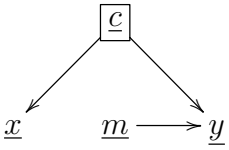


$= \sum_{x'} \sum_m P(y|m, x')P(x')P(m|x)$

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

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 $\rho_{\underline{m}}G$



(10.32)

QED

Chapter 11

D-Separation

Before reading this chapter, I recommend that you read Chapter 0.2 on the definition of bnets.

A path γ that isn't a loop can have 3 types of intermediate nodes $\underline{x}$ (an intermediate node of γ is a node in γ that isn't one of the two end nodes). Suppose $\underline{a}$ and $\underline{b}$ are the two neighbors of $\underline{x}$. Then the 3 possible cases are:

1. **mediator node:** $(\underline{a} \leftarrow \underline{x} \leftarrow \underline{b})$ or $(\underline{a} \rightarrow \underline{x} \rightarrow \underline{b})$
2. **fork node:** $(\underline{a} \leftarrow \underline{x} \rightarrow \underline{b})$
3. **collider node:** $(\underline{a} \rightarrow \underline{x} \leftarrow \underline{b})$

We say that a non-loop path γ from $\underline{a}$ to $\underline{b}$ (i.e., with end nodes $\underline{a}, \underline{b}$) is **blocked** by a multinode $\underline{Z}.$ if one or more of the following statements is true:

1. There is a node $\underline{x} \in \underline{Z}.$ which is a mediator or a fork of γ.
2. γ contains a collider node $\underline{c}$ and $(\underline{c} \cup de(\underline{c})) \cap \underline{Z}. = \emptyset$ (i.e., neither $\underline{c}$ nor any of the descendants of $\underline{c}$ is contained in $\underline{Z}.$)

This definition of a blocked path is easy to remember if one thinks of the following analogy with pipes carrying a fluid. Think of path γ as if it were a pipe carrying a fluid. Think of the nodes of γ as junctions in the pipe. If $\underline{Z}.$ intersects γ at either a mediator or a fork junction, that blocks the pipe flow. A collider junction $\underline{c}$ is like a blackhole or a huge leak. Its presence blocks passage of the fluid as long as neither $\underline{c}$ nor any of the descendants of $\underline{c}$ are in $\underline{Z}.$. If, on the other hand, $\underline{c} \in \underline{Z}.$, or $\underline{c}' \in \underline{Z}.$ where $\underline{c}' \in de(\underline{c})$, then that acts as a complete (in the case of $\underline{c} \in \underline{Z}.$) or a partial (in the case of $\underline{c}' \in \underline{Z}.$) bridge across the blackhole.

See Fig.11.1 for some examples of paths that are blocked or not blocked by a multinode $\underline{Z}.$.

Given 3 disjoint multinodes $\underline{A}., \underline{B}., \underline{Z}.$ of a graph G, we say " $\underline{A}. \perp \underline{B}.|\underline{Z}.$ in G" or " $\underline{A}.$ **and** $\underline{B}.$ **are d-separated by** $\underline{Z}.$" iff there exists no path γ from $\underline{a} \in \underline{A}.$, to $\underline{b} \in \underline{B}.$ which is not blocked by $\underline{Z}.$.

The minimal Markov blanket (see Chapter 23) of a node $\underline{a}$ is the smallest multinode $\underline{Z}.$ such that $\underline{a} \perp \underline{b}|\underline{Z}.$ for all $\underline{b} \notin \underline{a} \cup \underline{Z}.$.

We are finally ready to state the d-separation theorem, without proof.

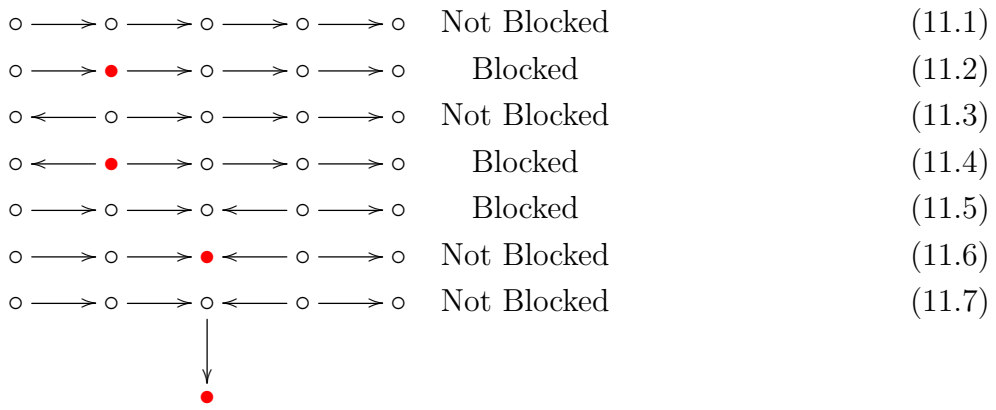


Figure 11.1: Examples of paths that are blocked or not blocked by a multinode $\underline{Z}.$. Nodes belonging to $\underline{Z}.$ are colored red.

Claim 5 *Suppose* $\underline{A}., \underline{B}., \underline{Z}.$ *are disjoint multinodes of a graph* G.
If $\underline{A}. \perp \underline{B}.|\underline{Z}.$ *in* G, *then* $P(B.|A., Z.) = P(B.|Z.)$.

Chapter 14

Front-door Adjustment

The front-door (FD) adjustment theorem is proven in Chapter 10 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 10 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 11.

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 adjustment 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 6 *Front-Door Adjustment Theorem*
If $\underline{m}.$ satisfies the front-door criterion relative to $(\underline{x}., \underline{y}.)$, and $P(x., m.) > 0$, then

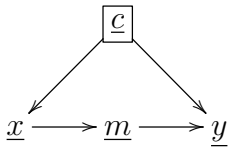
$$P(y.|\rho\underline{x}. = x.) = \sum_{m.} \underbrace{\left[\sum_{x.'} P(y.|x'., m.)P(x'.)\right]}_{P(y.|\rho\underline{m}.=m.)} \underbrace{P(m.|x.)}_{P(m.|\rho\underline{x}.=x.)} \tag{14.1}$$

$$= \sum_{m.,x.'} \left\{ \begin{array}{ccccc} & & \underline{x}. = x.' & & \\ & & & \searrow & \\ \underline{x}. = x. & \longrightarrow & \underline{m}. = m. & \longrightarrow & \underline{y}. \end{array} \right\} \tag{14.2}$$

proof: See Chapter 10
QED

Examples

1.



(14.3)

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

Chapter 22

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 Models (SEM)**. Judea Pearl in his books calls them **linear Structural Causal Models (SCM)**, because they are very convenient for doing causal analysis. We will refer to them as PA diagrams in honor of Sewall Wright.

A **PA diagram** is a special kind of bnet. To build a PA diagram, start with a linear deterministic bnet G. The deterministic nodes of G are called the **endogenous (internal) variables**. Now make a bigger bnet $\overline{G}$ called a PA diagram by adding to each node $\underline{a}$ of G a non-deterministic 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. They are assumed to be unobserved and their TPMs are prior probability distributions. Since they are root nodes, they are mutually independent. When we draw a PA diagram, we will never draw the exogenous nodes, leaving them implicit.

Example:

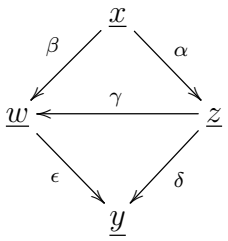


Figure 22.1: Example of a PA diagram wherein $\underline{x}$ splits into two nodes $\underline{z}$ and $\underline{w}$, then merges into node $\underline{y}$. There is also an arrow $\underline{z} \to \underline{w}$. Exogenous nodes are not shown. The Greek letters represent real numbers.

The TPMs, printed in blue, for the nodes of the PA diagram Fig.22.1, are as follows.

$$P(y|w,z,u_{\underline{y}}) = \mathbb{1}(y = \epsilon w + \delta z + u_{\underline{y}}) \tag{22.1}$$

$$P(w|x,z,u_{\underline{w}}) = \mathbb{1}(w = \beta x + \gamma z + u_{\underline{w}}) \tag{22.2}$$

$$P(z|x,u_{\underline{z}}) = \mathbb{1}(z = \alpha x + u_{\underline{z}}) \tag{22.3}$$

$$P(x|u_{\underline{x}}) = \mathbb{1}(x = u_{\underline{x}}) \tag{22.4}$$

Hence,

$$\begin{aligned} y &= \epsilon w + \delta z + u_{\underline{y}} && (22.5)\\ &= \epsilon(\beta x + \gamma z + u_{\underline{w}}) + \delta z + u_{\underline{y}} && (22.6)\\ &= (\epsilon\gamma + \delta)z + \epsilon\beta x + \epsilon u_{\underline{w}} + u_{\underline{y}} && (22.7)\\ &= (\epsilon\gamma + \delta)z + \epsilon\beta u_{\underline{x}} + \epsilon u_{\underline{w}} + u_{\underline{y}} \ . && (22.8) \end{aligned}$$

Therefore

$$\left(\frac{\partial y}{\partial z}\right)_{u_\cdot - u_{\underline{z}}} = \epsilon\gamma + \delta \ , \tag{22.9}$$

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 directed path from $\underline{z}$ to $\underline{y}(\underline{z})$. This is a general property of PA diagrams.

Fully Connected PA diagrams

The bnets that will be considered in this section will all be fully connected. Fully connected bnets are defined in Chapter 0.2. This section uses the notation $\langle \underline{x}, \underline{y} \rangle$ for the covariance of any two random variables $\underline{x}, \underline{y}$. This $\langle \underline{x}, \underline{y} \rangle$ notation is defined in the Notational Conventions Chapter 0.3.

Consider a PA diagram with deterministic nodes $\underline{x}_\cdot = (\underline{x}_k)_{k=0,1,\ldots nx-1}$ and corresponding exogenous nodes $\underline{u}_\cdot = (\underline{u}_k)_{k=0,1,\ldots nx-1}$. Assume $\langle \underline{u}_i, \underline{u}_j \rangle = 0$ if $i \neq j$. The strength of each connection $\underline{x}_i \to \underline{x}_j$ of the PA diagram is measured by a **structural coefficient** $\alpha_{j|i} \in \mathbb{R}$. Some of the $\alpha_{j|i}$ may be zero.

Fully connected PA diagram with $nx = 2$

Consider the PA diagram of Fig.22.2. This diagram represents the following **structural equations**:

$$\begin{aligned} \underline{x}_0 &= \underline{u}_0 && (22.10a)\\ \underline{x}_1 &= \alpha_{1|0}\underline{x}_0 + \underline{u}_1 \ . && (22.10b) \end{aligned}$$

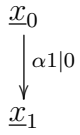


Figure 22.2: Fully connected PA diagram with two $\underline{x}_j$ nodes (exogenous nodes $\underline{u}_j$ not shown).

Eqs.22.10 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

$$\langle \underline{x}_1, \underline{x}_0 \rangle = \alpha_{1|0} \langle \underline{x}_0, \underline{x}_0 \rangle \ . \tag{22.11}$$

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

Fully connected PA diagram with $nx = 3$

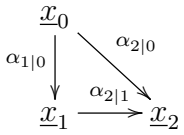


Figure 22.3: Fully connected PA diagram with three $\underline{x}_j$ nodes (exogenous nodes $\underline{u}_j$ not shown).

Consider the PA diagram of Fig.22.3. This diagram represents the following **structural equations**:

$$\underline{x}_0 = \underline{u}_0 \tag{22.12a}$$
$$\underline{x}_1 = \alpha_{1|0} \underline{x}_0 + \underline{u}_1 \tag{22.12b}$$
$$\underline{x}_2 = \alpha_{2|1} \underline{x}_1 + \alpha_{2|0} \underline{x}_0 + \underline{u}_2 \ . \tag{22.12c}$$

Eqs.22.12 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

$$\langle \underline{x}_1, \underline{x}_0 \rangle = \alpha_{1|0} \langle \underline{x}_0, \underline{x}_0 \rangle \tag{22.13a}$$
$$\langle \underline{x}_2, \underline{x}_0 \rangle = \alpha_{2|1} \langle \underline{x}_1, \underline{x}_0 \rangle + \alpha_{2|0} \langle \underline{x}_0, \underline{x}_0 \rangle \tag{22.13b}$$
$$\langle \underline{x}_2, \underline{x}_1 \rangle = \alpha_{2|1} \langle \underline{x}_1, \underline{x}_1 \rangle + \alpha_{2|0} \langle \underline{x}_0, \underline{x}_1 \rangle \tag{22.13c}$$

Eqs.22.13 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.

Fully connected PA diagram with arbitrary nx**.**

Let $\underline{x}. = (x_i)_{i=0,1,\ldots,nx-1}$ and $\underline{x}_{<i} = (x_k)_{k=0,1,\ldots,i-1}$. Consider a fully connected PA diagram with deterministic nodes labeled $\underline{x}_i$. The $\underline{x}_i$ labels are assumed to be in **topological order** (i.e.,

the parents of node $\underline{x}_i$ are $\underline{x}_{<i}$). Let the TPMs, printed in blue, for the nodes $\underline{x}.$ of the PA diagram, be

$$P(x_i|x_{<i}, u_i) = \mathbb{1}(x_i = \sum_{k<i} \alpha_{i|k} x_k + u_i) \ , \tag{22.14}$$

for some parameters $\alpha_{i|k} \in \mathbb{R}$. The exogenous nodes $\underline{u}.$ are assumed to be independent so

$$P(u.) = \prod_i P(u_i) \tag{22.15}$$

and

$$\langle \underline{u}_i, \underline{u}_j \rangle = 0 \text{ if } i \neq j \ . \tag{22.16}$$

Note that

$$\begin{aligned} P(x.) &= \sum_{u.} P(u.) \prod_i P(x_i|x_{<i}, u_i) & (22.17) \\ &= E_{\underline{u}.}[\prod_i P(x_i|x_{<i}, u_i)] \ . & (22.18) \end{aligned}$$

In terms of random variables, this system is described by the following **structural equations**:

$$\underline{x}_i = \sum_{k<i} \alpha_{i|k} \underline{x}_k + \underline{u}_i \ . \tag{22.19}$$

The structural equations can be written in matrix form as follows. Define a lower triangular matrix A with the connection strengths $\alpha_{i|k} \in \mathbb{R}$ as entries. For example, for $nx = 4$,

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \alpha_{1|0} & 0 & 0 & 0 \\ \alpha_{2|0} & \alpha_{2|1} & 0 & 0 \\ \alpha_{3|0} & \alpha_{3|1} & \alpha_{3|2} & 0 \end{bmatrix} \ . \tag{22.20}$$

If we now represent the multinodes $\underline{x}.$ and $\underline{u}.$ as column vectors $\underline{x}$ and $\underline{u}$, we get

$$\underline{x} = A\underline{x} + \underline{u} \ . \tag{22.21}$$

Note that

$$\underline{x} = (1 - A)^{-1} \underline{u} \ . \tag{22.22}$$

Therefore,

$$\underline{x}_i = f_i(\underline{u}_{\leq i}) \ . \tag{22.23}$$

Therefore, if $i > j$,

$$\langle \underline{u}_i, \underline{x}_j \rangle = \langle \underline{u}_i, f_j(\underline{u}_{\leq j}) \rangle = 0 \ . \tag{22.24}$$

Thus, if $i > j$,

$$\begin{aligned} \langle \underline{x}_i, \underline{x}_j \rangle &= \sum_{k<i} \alpha_{i|k} \langle \underline{x}_k, \underline{x}_j \rangle + \langle \underline{u}_i, \underline{x}_j \rangle & (22.25) \\ &= \sum_{k<i} \alpha_{i|k} \langle \underline{x}_k, \underline{x}_j \rangle & (22.26) \end{aligned}$$

Eqs.22.26 constitute a system of $(nx^2 - nx)/2$ linear equations in $(nx^2 - nx)/2$ 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] Wikipedia. Boolean algebra. https://en.wikipedia.org/wiki/Boolean_algebra.

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

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

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

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

[6] Christina Heinze-Deml. Causality, spring semester 2019 at ETH Zurich. https://stat.ethz.ch/lectures/ss19/causality.php#course_materials.

[7] 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/.

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

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

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

[11] 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.

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

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

[14] 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.

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

[16] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

[30] 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.

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

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

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

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

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

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

[37] 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".

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

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

[40] 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/.

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

[42] 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.

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