Chapter 3

Conditional Independence and Graphical Models

The first lecture dealt with the data matrix, geometry, distance between different objects. The second lecture dealt with principal component analysis, the aim of which is to reduce the dimension of the problem. Related to this was exploratory Factor Analysis, whereby the important principal components could suggest hidden (or latent) factors which influenced the variables.

We now turn to the topic of graphical models, where again we try to reduce the computational complexity of the problem. This time, we do this by finding and exploiting the independence structure between variables. We’ll also touch on causality; the attempt to ascertain whether one variable has a causal influence on other variables.

The idea of a graphical model is that the variables are represented as nodes on a graph and the edges in the graph represent a direct link between two variables. If two variables $X$ and $Y$ are graphically separated by a set of variables $S$, then all the influence that $X$ and $Y$ have on each other is mediated through the variables in $S$; $X$ and $Y$ are conditionally independent given $S$.

The problem of ascertaining statistical independence thereby becomes a problem of graphical separation and a powerful toolbox of graphical separation algorithms becomes available.

It is important to stress that, in a graphical model, graphical separation implies conditional independence, but the converse does not (in general) hold; there are often conditional independence relations that a graphical model cannot detect.

**Introduction** A graphical model for a probability distribution over several variables is, quite simply, a graph, where the random variables correspond to the node set of the graph and each graphical separation statement implies the corresponding conditional independence statement for the random variables. The opposite (that conditional independence implies graphical separation) in general does not hold. In a system with a large numbers of variables, the task of determining graphical separation statements is, in general, computationally far less demanding than the task of determining conditional independence.

A Bayesian network is the representation of a probability distribution on a directed acyclic graph
(DAG). In this setting, the most useful notion of separation is \textit{D-separation}, defined later. If a probability distribution factorises along a DAG, then D-separation statements in the DAG imply the corresponding conditional independence statements (although the reverse implication is, in general, false).

In many problems, for example gene expression data where there are thousands of variables, it may not be either possible or desirable to obtain a complete description of the dependence structure. The aim for such problems is to learn a DAG which encodes the most important features of the dependence structure. In classification problems, a complete description of the dependence structure is usually unnecessary; algorithms only locate the key features of the dependence structure to ensure accurate classification.

\section{Conditional Independence and Factorisation}

\textbf{Definition 3.1 (Independence).} Two random vectors $X$ and $Y$ are independent if their joint probability distribution factorises as

\[ P_{X,Y} = P_X P_Y. \]

$X$ and $Y$ are conditionally independent given a random vector $X$ if

\[ P_{X,Y|Z} = P_{X|Z} P_{Y|Z} P_Z. \]

This is written $X \perp Y|Z$.

\textbf{Example 3.1 (Binary Variables).}

Suppose $X_1, \ldots, X_d$ are binary variables (i.e. each takes values in \{0, 1\}). Then the state space is \{0, 1\}^d, which has $2^d$ possible configurations. To specify the probability distribution $P_{X_1, \ldots, X_d}$, we therefore need to specify $2^d - 1$ values (since all the values sum to 1).

Suppose that $X_1, \ldots, X_d$ are mutually independent. Then we only need to specify $d$ values; we need $P_{X_i}(1), \ldots, P_{X_i}(1)$, since $P_{X_i}(0) = 1 - P_{X_i}(1)$ for $i = 1, \ldots, d$.

For large $d$, there is therefore much computational advantage to be gained from exploiting the independence structure between the variables. \hfill \Box

\section{Definition of a Bayesian Network}

Consider a probability distribution over $d$ variables $P_{X_1, \ldots, X_d}$.

Recall that for any collection of events $A_1, \ldots, A_n$,

\[ P(A_1 \cap \ldots \cap A_n) = P(A_1) \frac{P(A_1 \cap A_2)}{P(A_1)} \ldots \frac{P(A_1 \cap \ldots \cap A_n)}{P(A_1 \cap \ldots \cap A_{n-1})} \]
so that (using the definition \( P(A|B) = \frac{P(A \cap B)}{P(B)} \)):

\[
P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1) \ldots P(A_n|A_1 \cap \ldots \cap A_{n-1}).
\]

Clearly, any probability distribution \( P_{X_1,\ldots,X_d} \) over \( \mathcal{X} \) may be factorised as

\[
P_{X_1,\ldots,X_d} = P_{X_{\sigma(1)}} \prod_{j=2}^{d} P_{X_{\sigma(j)|X_{\sigma(1)},\ldots,X_{\sigma(j-1)}}}
\]

for any permutation \( \sigma \) of 1, \ldots, \( d \). Let \( P_{a}^{(\sigma)}(j) \in \{ \sigma(1), \ldots, \sigma(j-1) \} \) satisfy

- \[
P_{X_1,\ldots,X_d} = P_{X_{\sigma(1)}} \prod_{j=2}^{d} P_{X_{\sigma(j)|P_{a}^{(\sigma)}(j)}}
\]

- \[
P_{X_1,\ldots,X_d} = P_{X_{\sigma(1)}} \prod_{j=2}^{d} P_{X_{\sigma(j)|\Theta(j)}}
\]

if any \( \Theta(j) \) is a *strict* subset of \( P_{a}^{(\sigma)} \).

Unless otherwise stated, it will be assumed that the variables are labelled in such a way that \( \sigma = I \),
the identity.

For \( P_{a} = \{ l_{j,1}, \ldots, l_{j,m_j} \} \), the state space of \( X_{\mathcal{P}_{a}(j)} \) is \( \mathcal{X}_{l_{j,1}} \times \ldots \times \mathcal{X}_{l_{j,m_j}} \). For discrete variables, there
are \( q_j = \prod_{a=1}^{m_j} k_{l_{j,a}} \) configurations. These may be labelled \( (\pi_{j,l_{j,a}}^{(l)})_{l_{j,a}} \) and the parameters required for the
probability distribution \( P_{X_1,\ldots,X_d} \) are

\[
\theta_{jil} = P_{X_{j|X_{\mathcal{P}_{a}(j)}}(i|\pi_{j,l_{j,a}}^{(l)})} \quad j = 1, \ldots, d \quad i = 0, \ldots, k_{j-1}, \quad l = 1, \ldots, q_j.
\]

**Estimating Parameters** Suppose we have an \( n \times d \) data matrix \( \mathbf{x} \); to estimate the parameter \( \theta_{jil} \),
we use

\[
\hat{\theta}_{jil} = \frac{\text{number of appearances of } (i, \pi_{j,l_{j,a}}^{(l)}) \text{ configuration in } \mathbf{x}}{\text{number of appearances of } \pi_{j,l_{j,a}}^{(l)} \text{ configuration in } \mathbf{x}}.
\]

**Factorising a Probability Distribution along a Directed Acyclic Graph** The factorisation of
a Bayesian network may be represented by a *Directed Acyclic Graph*. For example, if the probability
distribution over \( X, Y, Z, W \) satisfies

\[
P_{X,Y,Z,W} = P_{X}P_{Y|X}P_{Z|X}P_{W|Y,Z},
\]

the factorisation may be represented by the graph in Figure 3.1.
Figure 3.1: DAG representing the factorisation of a probability distribution

![DAG](image)

Figure 3.2: A Chain Connection

### 3.3 Connections in a Directed Acyclic Graph and Conditional Independence

**Definition 3.2 (Instantiated).** When the state of variable is known, the variable is said to be instantiated.

Within a directed acyclic graph, there are three basic ways in which two nodes $\alpha, \gamma$ such that $\alpha \rightarrow \gamma \not\in D$ and $\gamma \rightarrow \alpha \not\in D$ can be connected via a third node. They are the *chain*, *fork* and *collider* connections respectively.

**Chain Connections** A chain connection between nodes $\alpha$ and $\gamma$ is a connection via a node $\beta$ such that the graph contains directed edges $\alpha \rightarrow \beta$ and $\beta \rightarrow \gamma$, but no edge between $\alpha$ and $\gamma$.

Consider a probability distribution over $(X_\alpha, X_\beta, X_\gamma)$ factorised according to the graph in Figure 3.2, as $P_{X_\alpha}P_{X_\beta|X_\alpha}P_{X_\gamma|X_\beta}$.

Clearly, $X_\alpha \perp X_\gamma$ in general;

\[
P_{X_\alpha,X_\gamma}(x_1, x_3) = P_{X_\alpha}(x_1) \sum_{x_2 \in \mathbb{X}_2} P_{X_\beta|X_\alpha}(x_2|x_1) P_{X_\gamma|X_\beta}(x_3|x_2)
\]

and, without further assumptions, this cannot be expressed in product form.

Conditioned on the instantiation $X_\beta = x_2$,
3.3. Connections in a Directed Acyclic Graph and Conditional Independence

\[
P_{X_{\alpha}, X_{\beta}, X_{\gamma}} = \frac{P_{X_{\alpha} \mid X_{\beta}, X_{\gamma}}(\cdot \mid x_{2})}{P_{X_{\beta}}(x_{2})} = \frac{P_{X_{\alpha} \mid X_{\beta}}(x_{2}) P_{X_{\gamma} \mid X_{\beta}}(\cdot \mid x_{2})}{P_{X_{\beta}}(x_{2})}
\]

where Bayes rule has been used and so \( X_{\alpha} \perp X_{\gamma} \mid X_{\beta} \).

**Fork Connections**  A fork connection between two nodes \( X_{\alpha} \) and \( X_{\gamma} \) is a situation where there is no edge between \( X_{\alpha} \) and \( X_{\gamma} \), but there is a node \( X_{\beta} \) such that the graph contains directed edges \( X_{\beta} \rightarrow X_{\alpha} \) and \( X_{\beta} \rightarrow X_{\gamma} \). It is illustrated in Figure 3.3.

A distribution over the variables \((X_{\alpha}, X_{\beta}, X_{\gamma})\) that factorises according to the DAG in Figure 3.3 has factorisation

\[
P_{X_{\alpha}, X_{\beta}, X_{\gamma}} = P_{X_{\beta}} P_{X_{\alpha} \mid X_{\beta}} P_{X_{\gamma} \mid X_{\beta}}.
\]

It is clear that \( X_{\alpha} \not\perp X_{\gamma} \) in general;

\[
P_{X_{\alpha}, X_{\gamma}}(x_{1}, x_{3}) = \sum_{x_{2} \in X_{2}} P_{X_{\beta}}(x_{2}) P_{X_{\alpha} \mid X_{\beta}}(x_{1} \mid x_{2}) P_{X_{\gamma} \mid X_{\beta}}(x_{3} \mid x_{2})
\]

and, without further assumptions, this cannot be expressed in product form. Conditioned on \( X_{\beta} \), though:

\[
P_{X_{\alpha}, X_{\gamma} \mid X_{\beta}} = \frac{P_{X_{\alpha}, X_{\beta}, X_{\gamma}}}{P_{X_{\beta}}} = \frac{P_{X_{\beta}} P_{X_{\alpha} \mid X_{\beta}} P_{X_{\gamma} \mid X_{\beta}}}{P_{X_{\beta}}} = P_{X_{\alpha} \mid X_{\beta}} P_{X_{\gamma} \mid X_{\beta}}.
\]

It follows that \( X_{\alpha} \perp X_{\gamma} \mid X_{\beta} \).

**Collider Connections**  A collider connection between two nodes \( \alpha \) and \( \gamma \) is a connection such that the graph does not contain an edge between \( \alpha \) and \( \gamma \), but there is a node \( \beta \) such that the graph contains directed edges \( \alpha \rightarrow \beta \) and \( \gamma \rightarrow \beta \). A collider connection is illustrated in Figure 3.4.

The factorisation of the distribution \( P_{X_{\alpha}, X_{\beta}, X_{\gamma}} \) corresponding to the DAG for the collider is

\[
P_{X_{\alpha}, X_{\beta}, X_{\gamma}} = P_{X_{\alpha}} P_{X_{\beta}} P_{X_{\gamma} \mid X_{\alpha}, X_{\beta}}.
\]
In general, $X_\alpha \not\perp X_\gamma | X_\beta$. But for each $(x, z) \in \mathcal{X}_\alpha \times \mathcal{X}_\gamma$,

$$
P_{X_\alpha, X_\gamma}(x, z) = \sum_{y \in \mathcal{X}_\beta} P_{X_\alpha}(x) P_{X_\gamma}(z) P_{X_\beta|X_\alpha, X_\gamma}(y|x, z)
= P_{X_\alpha}(x) P_{X_\gamma}(z) \sum_{y \in \mathcal{X}_\beta} P_{X_\beta|X_\alpha, X_\gamma}(y|x, z)
= P_{X_\alpha}(x) P_{X_\gamma}(z).
$$

so that $X_\alpha \perp X_\gamma$.

**A Causal Interpretation** So far, the discussion has considered sets of random variables where, based on the ordering of the variables, the parent set of a variable is a subset of those of a lower order. The representation of a probability distribution by factorising along a Directed Acyclic Graph may be particularly useful if there are cause to effect relations between the variables, the ancestors being the cause and the descendants the effect. For a causal model, the connections have the following interpretations:

**Fork Connection: Common cause** For the fork connection, illustrated by Figure 3.2, $X_\beta$ may be a *cause* that influences both $X_\alpha$ and $X_\gamma$ which are *effects*. The variables are only related through $X_\beta$. The situation is illustrated by the following example, taken from a cartoon by Albert Engström; ‘during a convivial discussion at the bar one evening, about the unhygienic nature of galoshes, one of the participants pipes up, “you have a very good point there. Every time I wake up wearing my galoshes, I have a sore head.”

Let $X_\alpha$ denote the state of the feet and $X_\gamma$ the state of the head. These two variables are related; $X_\alpha \not\perp X_\beta$. But there is a common cause; $X_2$, which denotes the activities of the previous evening. Once it is known that he has spent a convivial evening drinking, the state of the feet gives no further information about the state of the head; $X_\alpha \perp X_\gamma | X_\beta$.

**Chain Connection** This may similarly be understood as cause to effect. $X_\alpha$ influences $X_\beta$, which in turn influences $X_\gamma$, but there is no direct causal relationship between the values taken by $X_\alpha$ and those taken by $X_\gamma$. If $X_\beta$ is unknown, then $X_\alpha \not\perp X_\gamma$, but once the state of $X_\beta$ is established, $X_\alpha$ and $X_\gamma$ give no further information about each other; $X_\alpha \perp X_\gamma | X_\beta$. 

![Figure 3.4: A Collider Connection](image-url)
3.4. Separation within a DAG

Attention is now turned to trails within a DAG, and characterisation of those along which information can pass.

Definition 3.3 (S-Active Trail). Let $\mathcal{G} = (V, D)$ be a directed acyclic graph. Let $S \subset V$ and let $\alpha, \beta \in V \setminus S$. A trail $\tau$ between the two variables $\alpha$ and $\beta$ is said to be $S$-active if

1. Every collider node in $\tau$ is in $S$, or has a descendant in $S$ (that is, for each collider node $\alpha \in \tau$, there is a directed path $\alpha \rightarrow \beta_1 \rightarrow \ldots \rightarrow \beta_m \rightarrow \gamma$ for some $\gamma \in S$).

2. Every other node is outside $S$.

Definition 3.4 (Blocked Trail). A trail between $\alpha$ and $\beta$ that is not $S$-active is said to be blocked by $S$.

The following definition is basic; it will be seen that if a probability distribution factorises along a DAG $\mathcal{G}$ and two nodes $\alpha$ and $\beta$ are $D$-separated by $S$, then $X_\alpha \perp X_\beta | X_S$.

Definition 3.5 (D-separation). Let $\mathcal{G} = (V, D)$ be a directed acyclic graph, where $V = \{1, \ldots, d\}$. Let $S \subset V$. Two distinct nodes $\alpha$ and $\beta$ not in $S$ are $D$-separated by $S$ if all trails between $\alpha$ and $\beta$ are blocked by $S$.

Let $A$ and $B$ denote two sets of nodes. If every trail from any node in $A$ to any node in $B$ is blocked by $S$, then the sets $A$ and $B$ are said to be $D$-separated by $S$. This is written

$$A \perp B \mid_G S.$$  \hspace{1cm} (3.1)

The terminology $D$-separation is short for directed separation. The insertion of the letter ‘$D$’ points out that this is not the standard use of the term ‘separation’ found in graph theory.

Definition 3.6 (D-connected). If two nodes $\alpha$ and $\beta$ are not $D$-separated, they are said to be $D$-connected.
**Notation**  The notation $\alpha \notin \beta \mid G S$ denotes that $\alpha$ and $\beta$ are $D$-connected by $S$ in the DAG $G$. Here $\alpha$ and $\beta$ may refer to individual nodes or sets of nodes.

**Example 3.2.**

Consider the chain connection $\alpha \rightarrow \beta \rightarrow \gamma$ in the DAG in Figure 3.2 and the fork connection of Figure 3.3. For the chain connection of Figure 3.2, the $D$-separation statements are: $\alpha \perp \gamma \mid G \beta$ while $\alpha \notin \beta \mid G \phi \ (\phi$ denotes the empty set). For the DAG in Figure 3.3, $\alpha \perp \gamma \mid G \beta$ while $\alpha \notin \gamma \mid G \beta$. These correspond to the conditional independence statements derived for probability distributions that factorise along these graphs. For Figure 3.4, $\alpha \perp \gamma \mid G \phi$ while $\alpha \notin \gamma \mid G \beta$. Again, these statements correspond to the conditional independence statements that may be derived from the fact that a distribution factorises along the DAG of Figure 3.4. □

Let $MB(\alpha)$ denote the set of nodes which are either parents of $\alpha$ or children of $\alpha$ or a node which shares a common child with $\alpha$. Then $\alpha$ is $D$-separated from the rest of the network by $MB(\alpha)$. This set of nodes is known as the Markov blanket of the node $\alpha$.

**Definition 3.7 (Markov Blanket).** The Markov blanket of a node $\alpha$ in a DAG $G = (V, D)$, denote $MB(\alpha)$, is the set consisting of the parents of $\alpha$, the children of $\alpha$ and the nodes sharing a common child with $\alpha$.

### 3.4.1 Bayes Ball

The Bayes ball provides a convenient method for deciding whether or not two nodes are $D$-separated by a set $S$ in a DAG $G = (V, D)$. Variables are $D$-connected by a set $S$ if the Bayes ball can be passed between them employing the following rule. The nodes which are not in $S$ are depicted as unshaded; nodes in $S$ as shaded.

**Definition 3.8 (Instantiated Nodes).** Let $G = (V, D)$ be a directed acyclic graph. When considering statements $\alpha \perp \beta \mid G S$ and $\alpha \notin \beta \mid G S$, the nodes in $S$ are referred to as instantiated.

Consider the three types of connection in a DAG; chain, collider and fork.

- For the chain connection illustrated in Figure 3.2, the Bayes ball algorithm indicates that if node $\beta$ is instantiated, then the ball does not move from $\alpha$ to $\gamma$ through $\beta$. The communication in the trail is blocked. If the node is not instantiated, then communication is possible.

- For the fork connection illustrated in Figure 3.3, the algorithm states that if node $\beta$ is instantiated, then again communication between $\alpha$ and $\gamma$ is blocked. If the node is not instantiated, then communication is possible.

- For the collider connection illustrated in Figure 3.4, the Bayes ball algorithm states that the ball does move from $\alpha$ to $\gamma$ if node $\alpha$ or any of its descendants is instantiated. If $\beta$ or a descendant is instantiated, this opens communication between the parents. If neither $\beta$ nor any of its descendants are instantiated, then there is no communication.
For a collider node $\beta$, instantiating any of the descendants of $\beta$ also opens communication. If node $\beta$ is not instantiated, and none of its descendants are instantiated, then there is no communication.

A DAG $G = (V, D)$ satisfies the following important property:

**Theorem 3.9.** A DAG $G = (V, D)$ contains an edge between two nodes $\alpha, \beta \in V$ if and only if $\alpha \perp \beta | S$ for any $S \subseteq V \setminus \{\alpha, \beta\}$.

**Proof** The proof of this is straightforward and left as an exercise. \hfill $\Box$

### 3.5 D-Separation and Conditional Independence

The following key result shows that if a probability distribution factorises along a given DAG $G$, then every $D$-separation statement for the DAG implies the corresponding conditional independence statement for the distribution.

**Theorem 3.10 (D-Separation Implies Conditional Independence).** Let $G = (V, D)$ be a directed acyclic graph and let $\mathbb{P}$ be a probability distribution that factorises along $G$. Then for any three disjoint subsets $A, B, S \subset V$, it holds that $X_A \perp X_B | X_S$ ($X_A$ and $X_B$ are independent given $X_S$) if $A \perp B | g A$ ($A$ and $B$ are $D$-separated by $S$).

**Proof of Theorem 3.10** Omitted \hfill $\Box$

Of course, the converse is not true in general; $D$-separation is a convenient way of locating some of the independence structure of a distribution. It does not, in general, locate the entire independence structure.
3.6 Queries

Once a probability distribution has been factorised according to a Bayesian Network, the next task is to use it to answer queries.

**Definition 3.11 (Query).** A query in probabilistic inference is simply a conditional probability distribution, over the variables of interest (the query variables) conditioned on information received.

3.7 Bayesian Networks in R

3.8 Introduction

It has become clear that R is now the most effective and dominant language of statistical computing. There are excellent packages available in R for Bayesian Networks, for inference using a given Bayesian Network and for learning the structure of a Bayesian Network. This chapter introduces some of the software in R available for Bayesian Networks and discusses graphs in R and inference using networks that have already been defined. Parameter learning and structure learning are considered later.

The packages considered are **gRain** by Søren Højsgaard and **bnlearn**.

Having installed R and a suitable editor (for example Rstudio), the relevant packages have to be installed.

**gRain and related packages**  Information for **gRain** is available on the author’s web page:

http://people.math.aau.dk/~sorenh/software/gR/

The package, along with all the supporting packages, has to be installed. As pointed out on the web page, the package uses the packages **graph**, **RBGL** and **Rgraphviz**. These packages are *not* on CRAN, but on ‘bioconductor’. To install these packages, execute

```r
install.packages("BiocManager")
setRepositories()
```

and then make sure that all are activated (2 3 4 5 6 7 8)

Now install using:

```r
install.packages("gRbase", dependencies=TRUE);
install.packages("gRain", dependencies=TRUE);
install.packages("gRim", dependencies=TRUE)
```

The package **bnlearn** also has some useful inference functions, although its main consideration is learning. Install it in the usual way:

```r
> install.packages("bnlearn")
```
3.9 Graphs in R

This section considers the various graphs that appear in graphical modelling and how to render them in R. In addition to the packages mentioned so far, the package `ggm`, has some useful functions for graphical Markov models.

```r
> install.packages("ggm")
```

Another useful graphics package is `igraph`

```r
> install.packages("igraph")
```

We also need the package RBGL, which is not available on CRAN, but is only available on BioConductor. So set repositories with

```r
setRepositories()
```

make sure that the appropriate repositories are checked and then

```r
install.packages("RBGL")
```

These packages should be activated:

```r
> library("bnlearn")
> library("gRain")
> library("ggm")
> library("igraph")
> library("RBGL")
> library("RBase")
```

3.10 Example: ‘Asia’ by Lauritzen

We consider the ‘Asia’ example of Lauritzen et. al. You have returned from holiday in Asia and you are feeling unwell. There may be nothing seriously wrong with you, but you could be suffering from tuberculosis, lung cancer or bronchitis. The causal diagram is shown in Figure 3.6. Let $A$ denote ‘visit to Asia’ with values ‘yes’ or ‘no’. Similarly, all the other variables are binary and are labelled $S$ for smoker, $T$ for tuberculosis, $L$ for lung cancer, $B$ for bronchitis, $E$ for either, $X$ for X-ray (‘yes’ for indication of a problem, ‘no’ for clear), $D$ for dyspnoea (shortness of breath)

\[
\begin{align*}
\mathbb{P}(A = \text{yes}) &= 0.01 \\
\mathbb{P}(T = \text{yes}|A = \text{yes}) &= 0.05 \\
\mathbb{P}(T = \text{yes}|A = \text{no}) &= 0.01 \\
\mathbb{P}(S = \text{yes}) &= 0.5 
\end{align*}
\]
\[
\begin{align*}
\mathbb{P}(L = \text{yes} | S = \text{yes}) & = 0.1 & \mathbb{P}(L = \text{yes} | S = \text{no}) & = 0.01 \\
\mathbb{P}(B = \text{yes} | S = \text{yes}) & = 0.6 & \mathbb{P}(B = \text{yes} | S = \text{no}) & = 0.3 \\
\mathbb{P}(E = \text{yes} | L = \text{yes}, B = \text{yes}) & = 1, & \mathbb{P}(E = \text{yes} | L = \text{yes}, B = \text{no}) & = 1 \\
\mathbb{P}(E = \text{yes} | L = \text{no}, B = \text{yes}) & = 1, & \mathbb{P}(E = \text{yes} | L = \text{no}, B = \text{no}) & = 0 \\
\mathbb{P}(X = \text{yes} | E = \text{yes}) & = 0.98 & \mathbb{P}(X = \text{yes} | E = \text{no}) & = 0.05 \\
\mathbb{P}(D = \text{yes} | B = \text{yes}, E = \text{yes}) & = 0.9 & \mathbb{P}(D = \text{yes} | B = \text{yes}, E = \text{no}) & = 0.7 \\
\mathbb{P}(D = \text{yes} | B = \text{no}, E = \text{yes}) & = 0.8 & \mathbb{P}(D = \text{yes} | B = \text{no}, E = \text{no}) & = 0.1
\end{align*}
\]

We can programme the network into R as follows. We need the packages \texttt{gRain} and \texttt{gRbase}. The conditional probability potentials may be specified as follows:

```r
> library("gRain")
Loading required package: gRbase
> yn <- c("yes","no")
> a <- cptable(~asia, values=c(1,99), levels=yn)
> t.a <- cptable(~tub+asia, values=c(5,95,1,99), levels=yn)
> s <- cptable(~smoke, values=c(5,5), levels=yn)
> l.s <- cptable(~lung+smoke, values=c(1,9,1,99), levels=yn)
> b.s <- cptable(~brcm+smoke, values=c(6,4,3,7), levels=yn)
> e.lt <- cptable(~either+lung+tub, values=c(1,0,1,0,1,0,1), levels=yn)
> x.e <- cptable(~xray+either, values=c(98,2,5,95), levels=yn)
> d.be <- cptable(~dyasp+brcm+either, values=c(9,1,7,3,8,2,1,9), levels = yn)
```

The + operator could be considered slightly misleading. There are other ways to enter the conditional probability potentials:

```r
> t.a <- cptable(~tub|asia, values=c(5,95,1,99), levels=yn)
> t.a <- cptable(~"tub","asia"), values=c(5,95,1,99), levels=yn)
```

There are also special functions \texttt{ortable()} and \texttt{andtable}. For example, \texttt{e.lt()} could be entered by:

```r
> e.lt <- ortable(~either+lung+tub, levels=yn)
```

### 3.10.1 Building the Network

A network is created with the function \texttt{grain()}, which returns an object of class \texttt{grain}:
3.10. EXAMPLE: ‘ASIA’ BY LAURITZEN

> plist<-compileCPT(list(a, t.a, s, l.s, b.s, e.lt, x.e, d.be))
> grn1<-grain(plist)
> summary(grn1)

Independence network: Compiled: FALSE Propagated: FALSE
Nodes : chr [1:8] "asia" "tub" "smoke" "lung" "bronc" "either" ...
> plot(grn1)

![Graph of Asia Network]

Figure 3.6: Asia Network

The plot is shown in Figure 3.6.

3.10.2 Compilation

The network has to be compiled and propagated before queries can be made.

> grn1c<-compile(grn1)
> summary(grn1c)

Independence network: Compiled: TRUE Propagated: FALSE
Nodes : chr [1:8] "asia" "tub" "smoke" "lung" "bronc" "either" ...
Number of cliques: 6
Maximal clique size: 3
Maximal state space in cliques: 8

3.10.3 Absorbing Evidence and Answering Queries

Evidence may be entered as follows: for example, suppose we have evidence that someone has visited asia and has dyspnoea. This is entered as follows:

> grn1c.ev<-
+ setFinding(grn1c,nodes=c("asia","dysp"),states=c("yes","yes"))
This creates a new grain object. The grain objects with (grnic.ev) and without (grnic) can be queried to give marginal probabilities:

```r
> querygrain(grnic.ev, nodes=c("lung","bronc"), type="marginal")
$lung
lung
   yes  no
0.09952515 0.90047485

$bronc
bronc
   yes  no
0.8114021  0.1885979
```

```r
> querygrain(grnic, nodes=c("lung","bronc"), type="marginal")
$lung
lung
   yes  no
0.055  0.945

$bronc
bronc
   yes  no
0.45  0.55
```

The evidence in a grain object can be retrieved with the `getFinding()` function, while the probability of observing the evidence is obtained using the `pFinding()` function:

```r
> getFinding(grnic.ev)
Finding:
asia: yes
dysp: yes
Pr(Finding)= 0.004501375
> pFinding(grnic.ev)
[1] 0.004501375
```

Joint and conditional distributions may be computed as follows:

```r
> querygrain(grnic.ev, nodes=c("lung","bronc"), type="joint")
bronc
lung     yes    no
   yes 0.06298076 0.03654439
```
no 0.74042132 0.15205354
> querygrain(grn1c.ev,nodes=c("lung","brnc"),type="conditional")
  bronc
      lung   yes     no
    yes 0.07761966 0.1937688
     no 0.92238034 0.8062312

These are both conditioned on the evidence; the former the joint distribution of lung and bronc conditioned on the evidence, while the latter is the conditional distribution of lung given bronc and the evidence.
Chapter 4

Intervention Calculus

4.1 Causal Models and Bayesian Networks

In many applications, a Bayesian network is constructed as a causal model, where for each variable, its parent variables are considered to be direct causes that influence the value taken by the variable.

For example, an earth tremor or a burglary can cause the burglar alarm to go off and the arrows in the associated collider DAG represent cause to effect relations. It is self-evident, but nevertheless has to be stated, that only associations can be inferred from an \( n \times d \) data matrix \( x \) of instantiations; directions of cause to effect cannot be inferred from data alone. When conditional independence statements are learned from data, this can be interpreted as a Markov model and it may be possible to construct an efficient factorisation of the distribution using these conditional independence statements. Clearly, this factorisation cannot be understood as a causal model, unless there are other modelling assumptions. For example, consider a model containing observable variables \( A, B, C \), where there are hidden variables \( H_1, H_2 \) that are unknown to the experimenter. If the causal diagram representing the causal relations between these variables is given by the DAG on the left in Figure 4.1, then the learned DAG, along which the distribution of \( A, B, C \) can be factorised, is the DAG on the right of Figure 4.1.

This is the correct DAG, in that it preserves the \( d \)-connection properties between \( A, B, C \), but the collider connection cannot be interpreted as \( A \) and \( B \) having a causal effect on \( C \); they are effects of the latent common causes \( H_1 \) and \( H_2 \).

If a Bayesian network is to be interpreted as a causal model, then the possible directions of cause

![Figure 4.1: Hidden causes and the learned DAG](image-url)
to effect must be part of the modelling assumptions before the data is analysed, determined by other considerations. The data analysis only determines which directed edges remain and which are removed. From data, one can determine whether or not there is an association between earth tremors and alarms triggered; it is not possible to determine from the data what causes what.

This is self evident, but surprisingly it turns out that it is necessary to state this. An article by Freedman and Humphreys from 1999 pointed out the obvious fact that causality could not be inferred from data alone and was a necessary response to obvious errors in the literature, where the term 'causal discovery' has been used in surprising ways, even after it had been established, with simple concrete and obvious examples, that the concept was ridiculous and long after publication of the Friedman Humphreys article illustrating that it was ridiculous. The article by Freedman and Humphreys is a good article; it is surprising that the literature had degenerated to such an extent that it was necessary for the authors to write it.

To define a causal network, an additional ingredient is needed; this is the concept of intervention, introduced by Judea Pearl in a seminal article from 1995.

### 4.2 Conditioning by Observation and by Intervention

Let $X$ and $Y$ be two random variables and suppose that $X = x$ is observed. Then the conditional probability of $Y = y$ is defined as

$$P_{Y|X}(y|x) = \frac{P_{X,Y}(x,y)}{P_X(x)}.$$  

This formula describes the way that the probability distribution of the random variable $Y$ changes after $X = x$ is observed. If, instead, the value $X = x$ is forced by the observer, irrespective of other considerations, the conditional probability statement is invalid.

If random variables are linked through a causal model, expressed by a directed acyclic graph, where parent variables have a causal effect on their children, some attempt can be made to compute the probability distribution over the remaining variables when the states of some variables are forced.

In a controlled experiment, a variable is forced to take a particular value, chosen at random, irrespective of the other variables in the network. In terms of the directed acyclic graph, the variable is instantiated with this value, the directed edges between the variable and its parents are removed (because the parents no longer have influence on the state of the variable) and all other conditional probabilities remain unaltered.

### 4.3 The Intervention Calculus for a Bayesian Network

**Definition 4.1** (The Intervention Formula). The conditional probability of $X_{V \setminus A} = x_{V \setminus A}$, given that the variables $X_A$ were forced to take the values $x_A$ independently of all else, is written

$$P_{V \setminus A} (x_{V \setminus A} | X_A \leftarrow x_A) \quad \text{or} \quad P_{V \setminus A} (x_{V} | x_{A})$$
4.3. THE INTERVENTION CALCULUS FOR A BAYESIAN NETWORK

and defined as

\[ \mathbb{P}_{V \setminus A | A}(x_{V \setminus A} | x_A) = \mathbb{P}_{V \setminus A | A}(x_{V \setminus A} | x_A) = \prod_{v \in V \setminus A} \mathbb{P}_{v | Pa(v)}(x_v | x Pa(v)). \tag{4.1} \]

Note that (4.1) is equivalent to:

\[ \mathbb{P}_{V \setminus A | A}(x_{V \setminus A} | x_A) = \frac{\mathbb{P}_V(x_V)}{\prod_{v \in A} \mathbb{P}_{v | Pa(v)}(x_v | x Pa(v))}. \tag{4.2} \]

The last expression of Equation (4.1) is in terms of the required factorisation; instantiation of the variables indexed by the set A and elimination of those edges in D which lead from the parents of the nodes in A to the nodes in V \setminus A. The terminology ‘local surgery’ is used to describe such an elimination.

A local surgery is performed and the conditional probabilities on the remaining edges are multiplied. This yields a factorisation along a mutilated graph where the direct causes of the manipulated variable are put out of effect.

The intervention formula (4.1) is obtained by wiping out those factors from the factorisation which correspond to the interventions. An explicit translation of intervention in terms of ‘wiping out’ equations was first proposed by Strotz and Wold (1960).

The quantity \( \mathbb{P}_{V \setminus A | A}(x_{V \setminus A}) \) from Definition 4.1 defines a family of probability measures over \( X_{V \setminus A} \), which depends on the values \( x_A \), which may be considered as parameters. These are the values forced on the variables indexed by A. This family includes original probability measure; if \( A = \phi \), then \( \mathbb{P}_{V \setminus A | A}(x_{V \setminus A}) = \mathbb{P}_X(.) \). This family is known as the intervention measure. In addition, the expression on the right hand side of (4.1) is called the intervention formula.

**Intervention** An ‘intervention’ is an action taken to force a variable into a certain state, without reference to its own current state, or the states of any of the other variables. It may be thought of as choosing the values \( x_A^* \) for the variables \( X_A \) by using a random generator independent of the variables \( X \).

**Remark** In the same style of notation, conditioning by observation is

\[ \mathbb{P}_{X \setminus A | x_A}(x_{V \setminus A} | \text{see}(x_A)) = \mathbb{P}_{X \setminus A | x_A}(x_{V \setminus A} | x_A) \tag{4.3} \]

where, by the standard definition of conditional probability,

\[ \mathbb{P}_{V \setminus A | A}(x_{V \setminus A} | x_A) = \frac{\mathbb{P}_V(x)}{\mathbb{P}_X(x_A)}. \tag{4.4} \]

**Example 4.1.**

Consider the DAG given in Figure 4.2, for ‘X having causal effect on Y’.

The factorisation of \( \mathbb{P}_{X,Y} \) along the DAG in Figure 4.2 is

\[ \mathbb{P}_{X,Y}(x,y) = \mathbb{P}_{Y|X}(y|x) \mathbb{P}_X(x) \]
and the intervention formula gives
\[
P_{Y|X}(y|x) = P_{Y|X}(y|x).
\]
Since \(X\) is a parent of \(Y\), the intervention to force \(X = x\) produces exactly the same conditional probability distribution over \(Y\) as observing \(X = x\). But if instead \(Y\) is forced, the intervention formula yields
\[
P_{X|Y}(x|y) = P_X(x).
\]
Clearly, \(P_{X|Y}(x|y) \neq P_{X|Y}(x|y)\) as functions unless \(X\) and \(Y\) are independent.

**Example 4.2** (The DAG for a wet pavement).

The ‘wet pavement’ example is a classic illustration, introduced by Judea Pearl. The DAG represents a causal model for a wet pavement and is given in Figure 4.3. The season \(A\) has four states; spring, summer, autumn, winter. Rain \(B\) has two states; yes / no. Sprinkler \(C\) has two states; on / off. Wet pavement \(D\) has two states; yes / no. Slippery pavement \(E\) has two states; yes / no.

The joint probability distribution is factorised as
\[
P_{A,B,C,D,E} = P_AP_B|A^C|A^D|B,C^E|D.
\]
Suppose, without reference to the values of any of the other variables and without reference to the current state of the sprinkler, ‘sprinkler on’ is now enforced. This could be, for example, regular maintenance work, which is carried out at regular intervals, irrespective of the season or other considerations. Then
\[ P_{A,B,D,E|C}(\cdot|C \leftarrow 1) = P_A P_B|A P_D|B,C(\cdot,1) P_E|D. \]

After observing that the sprinkler is on, it may be inferred that the season is dry and that it probably did not rain and so on. If ‘sprinkler on’ is enforced, without reference to the state of the system when the action is taken, then no such inference should be drawn in evaluating the effects of the intervention. The resulting DAG is given in Figure 4.4. It is the same as before, except that \( C = 1 \) is fixed and the edge between \( C \) and \( A \) disappears. The deletion of the factor \( P_{C|A} \) represents the understanding that whatever relationships existed between sprinklers and seasons prior to the action, found from

\[ P_{A,B,D,E|C}(\cdot,\cdot,\cdot,1) \]

are no longer in effect when the state of the variable is forced, as in a controlled experiment, without reference to the state of the system.

![DAG diagram](image)

Figure 4.4: Sprinkler ‘on’ is forced

After observing that the sprinkler is on, it may be inferred that the season is dry, that it probably did not rain and so on. No such inferences may be drawn in evaluating the effects of the intervention ‘ensure that the sprinkler is on’.

\[ \square \]

4.4 Causal Models

Having defined the family of intervention measures, the concept of causal model may now be defined.

**Definition 4.2** (Causal Model). Let \( X = (X_1, \ldots, X_d) \) be a random vector and let \( V = \{1, \ldots, d\} \) denote the indexing set. A causal model consists of the following:

1. A Bayesian Network for \( P_X \), that is, an ordering \( \sigma \) of the indices \( V \), a factorisation of the probability distribution

\[ P_V = \prod_{j=1}^{d} P_{\sigma(j)|P_{d^{\sigma}(j)}}(\sigma(j)) \quad (4.5) \]

where \( P_{\sigma^{\sigma}(j)} \subseteq \{\sigma(1), \ldots, \sigma(j-1)\} \) and is the smallest such subset such that (4.5) holds.
2. The node set $V$ consists of two types of nodes; $V_I$ and $V_N$, where $V_I \cap V_N = \emptyset$ and $V_I \cup V_N = V$. The nodes $V_I$ are the interventional nodes and $V_I$ are the non-interventional nodes, where no intervention is possible. The intervention formula (4.1) holds for each subset $A \subseteq V_I$ of interventional nodes and each $x_A \in A$.

The arrows $\alpha \rightarrow \beta$ of the DAG for either $\alpha$ or $\beta$ (or both) in $V_I$ are causal arrows, indicating direct cause to effect. The remaining arrows are non-causal; a cause to effect relation between nodes $\alpha$ and $\beta$ cannot be inferred from an arrow $\alpha \rightarrow \beta$ if both $\alpha, \beta \in V_N$.

### 4.4.1 Establishing a Causal Model via a Controlled Experiment

If sufficient data is available, a suitable Bayesian Network may be learned from the data. A causal model cannot be established from data alone. Additional information is needed, which is obtained through interventions on the interventional variables.

For example, the three graphs in Figure 4.5 are Markov equivalent; if the probability distribution factorises along one of these graphs, it also factorises along the others. The chains $\alpha \rightarrow \gamma \rightarrow \beta$ and $\alpha \leftarrow \gamma \leftarrow \beta$ and the fork $\alpha \leftarrow \gamma \rightarrow \beta$ are all Markov equivalent, with $D$-separation structure $\alpha \perp \beta | \gamma$. If any of these DAGs represents a causal network, then it is not possible to learn the causal network from the data alone.

Suppose that it is possible to intervene by controlling the variable $X_\gamma$, then if one of these graphs is the DAG for a causal network, it will be possible to establish which one through a controlled experiment. Figure 4.6 shows the associated structural model when the control $X_\gamma \leftarrow z$ has been applied, forcing $X_\gamma$ to be independent of its ancestors. A controlled experiment, where the direct causal links between $X_\gamma$ and its parent variables have been eliminated, will exhibit independence structure $X_\alpha \perp \{X_\beta, X_\gamma\}$ in the first case, $X_\alpha \perp X_\beta | X_\gamma$ in the second $\{X_\alpha, X_\gamma\} \perp X_\beta$ in the third. Once the associations $X_\alpha \perp X_\beta | X_\gamma$, $X_\alpha \not\perp X_\beta$, $X_\alpha \not\perp X_\gamma$, $X_\alpha \not\perp X_\beta \not\perp X_\gamma$, $X_\beta \not\perp X_\gamma$ and $X_\beta \not\perp X_\gamma | X_\alpha$ have been established, an additional controlled experiment, if it is possible to control the variable $X_\gamma$ with interventions to force all possible values of $X_\gamma$, will determine which graph within the equivalence class is appropriate.

![Figure 4.5: Three Markov Equivalent Graphs](image)

If it is possible to control variables, then it is possible to learn whether or not a collider represents independent causes with a common effect. If the DAG on the left hand side of Figure 4.1 represents a causal structure, then an experiment where variable $A$ is controlled will establish that it is not a direct cause of $C$, since an intervention on $A$ leaves it separated from the rest of the network, as in Figure 4.7.
4.5. Confounding, The ‘Sure Thing’ Principle and Simpson’s Paradox

4.5.1 Confounding

Consider the DAG given in Figure 4.8. It corresponds to the factorisation:

\[
P_{A,B,C} = P_{B|A,C}P_{A|C}P_C.
\]

Consider the conditional probability of \(B\), when \(A\) is controlled; \(P_{B|A}(\cdot | a)\). The DAG illustrating the intervention is shown in Figure 4.9. Note that

\[
P_{B|A}(\cdot | a) = \sum_{c \in C} P_{B,C|A}(\cdot , c|a).
\]

and that

\[
P_{B,C|A}(\cdot , | a) = P_{B|C|A}(\cdot , | a)P_{C|A}(\cdot | a) = P_{B|A,C}(\cdot , | a)P_C.
\]

Figure 4.6: Graphs from Figure 4.5 with intervention \(X_\gamma \leftarrow z\) applied

Figure 4.7: Hidden causes \(H_1\) and \(H_2\); intervention \(A = a\)

Figure 4.8: Illustration for Confounding
where in the second term, the do-conditioning of $A \leftarrow a$ is applied first, and then $C$ is observed. It follows that

$$P_{B|A}(\cdot|a) = \sum_{c \in \mathcal{C}} P_{B|A,C}(\cdot|a,c)P_C(c).$$

This shows that to estimate $P_{B|A}(\cdot|a)$ from data alone (i.e. without controlling $A$), it is necessary to be able to estimate $P_{B|A,C}$ and $P_C$ from data. If $C$ is observable, then the effect on the probability distribution of $B$ of manipulating $A$ may be estimated. But if $C$ is a hidden random variable (sometimes the term latent is used) in the sense that no direct sample of the outcomes of $C$ may be obtained, it will not be possible to estimate the probabilities used on the right hand side and hence it will not be possible to predict the effect on $B$ of manipulating $A$. This is known as confounding.

### 4.5.2 Simpson’s Paradox

Consider three binary variables, $A, B$ and $C$. Simpson’s paradox is the observation that there are situations where

$$\frac{P_{B|C,A}(1|1,1)/P_{B|C,A}(0|1,1)}{P_{B|C,A}(1|1,0)/P_{B|C,A}(0|1,0)} > 1 \quad \text{and} \quad \frac{P_{B|C,A}(1|0,1)/P_{B|C,A}(0|0,1)}{P_{B|C,A}(1|0,0)/P_{B|C,A}(0|0,0)} > 1,$$

but

$$\frac{P_{B|A}(1|1)/P_{B|A}(0|1)}{P_{B|A}(1|0)/P_{B|A}(0|0)} < 1.$$

For example let $A$ denote ‘treatment’, $B$ ‘recovery’ and $C$ ‘blood pressure’. Simpson’s paradox states that even if the ‘treatment’ may improve the chances of recovery for those with high blood pressure and those with low blood pressure, it may nevertheless be bad for the population as a whole. It could be that although the treatment is comparatively good within the group where high blood pressure is observed after treatment and also comparatively good within the group where low blood pressure is observed after treatment, it may be bad for the population as a whole. This occurs if ‘treatment’ increases blood pressure and increased blood pressure reduces the chances of recovery.

This situation is illustrated by the DAG given in Figure 4.10, where $A$ denotes treatment, $B$ recovery and $C$ blood pressure. Suppose that $C$ is a hidden variable. Even if the ‘treatment’ variable
A can be controlled, an intervention on $A$ does not remove any arrows from the causal diagram; there is the possibility of a Simpson’s paradox, even with a controlled experiment.

If $A$ denotes ‘treatment’ and $B$ ‘recovery’ and $C$ denotes a common cause of both $A$ and $B$, as in Figure 4.8, Simpson’s paradox may be resolved if $A$ can be controlled, because controlling $A$ breaks the causal link between $C$ and $A$. This is the sure thing principle, considered next, which states that if the treatment improves the chances of recovery for each level of the ‘common cause’ variable $C$, then it is good for the population as a whole.

![Causal Diagram](image)

**Figure 4.10:** $A=$treatment / $B=$recovery / $C=$blood pressure

### 4.5.3 The Sure Thing Principle

Consider again the situation of Figure 4.8. Suppose that $A$ is controlled; values for the variable $A$ are assigned at random, so the link $C \rightarrow A$ is broken and hence the effect on $B$ of manipulating $A$ is not confounded by the effects of hidden variables. The following result is referred to as ‘The Sure Thing Principle’. It states that when Figure 4.8 represents the causal structure and there is do-conditioning on $A$, then Simpson’s paradox does not hold.

**Proposition 4.3.** Consider three binary variables $A$, $B$, $C$ with the network given in Figure 4.8. If

$$\Pr_{B|C|A}(1|1|1) < \Pr_{B|C|A}(1|1|0)$$

and

$$\Pr_{B|C|A}(1|0|1) < \Pr_{B|C|A}(1|0|0)$$

then

$$\Pr_{B|A}(1|1|1) < \Pr_{B|A}(1|0|0).$$

The notation means: first $A$ is forced, then $C$ is observed.

**Proof**  Firstly,

$$\Pr_{B|A}(1|1|1) = \Pr_{B|C|A}(1|1|1)\Pr_{C|A}(1|1|1) + \Pr_{B|C|A}(1|0|1)\Pr_{C|A}(0|1).$$

Since $C$ is a parent of $A$,
\[ P_{C|A}(\|1) = P_C(\cdot). \]

It follows that
\[
P_{B|A}(1|1) = \sum_{x=0}^{1} P_{B|C,A}(1|x|1)P_{C|A}(x|1) = \sum_{x=0}^{1} P_{B|C,A}(1|x|1)P_{C}(x).
\]

Similarly,
\[
P_{B|A}(1|0) = \sum_{x=0}^{1} P_{B|C,A}(1|x|0)P_{C}(x).
\]

It now follows directly from the assumptions that
\[
P_{B|A}(1|1) < P_{B|A}(1|0),
\]
which is the stated result. \( \square \)

**Identifiability** An effect is said to be *identifiable* if it can be estimated from data alone.

In the example above, when \( C \) is a *common cause* of both \( A \) and \( B \), the *effect* of \( A \) on \( B \) is *identifiable*.

When \( C \) is on the *causal path* between \( A \) and \( B \), the effect of treatment \( A \) on condition \( B \) is *not* identifiable from data alone; it is necessary to additionally *control* for the effect of \( B \).

### 4.6 Identifiability: Back-Door and Front-Door Criteria

In a wide variety of situations, the aim is to compute the effects of an intervention, when it is not possible to carry out a controlled experiment. The following example, introduced by Pearl, introduced the issues involved.

**Example 4.3.**

Consider an experiment in which soil fumigants \( X \) are to be used to increase oat crop yields \( Y \), by controlling the eelworm population, \( Z \). These may also have direct effects, both beneficial and adverse, on yields, besides the control of eelworms. We would like to assess the total effects of the fumigants on yields when the study is complicated by several factors. First, controlled, randomised experiments are infeasible: farmers insist on deciding for themselves which plots are to be fumigated. Secondly, the farmers’ choice of treatment depends on last year’s eelworm population \( Z_0 \). This is an unknown quantity, but is strongly correlated with this year’s population. This presents a classic case of confounding bias, which interferes with the assessment of the treatment effects, regardless of sample size. Fortunately, through laboratory analysis of soil samples, the eelworm populations before and after treatment can be determined. Furthermore, since fumigants are only active for a short period, they do not affect the growth of eelworms surviving the treatment; eelworm growth depends on the
4.6. IDENTIFIABILITY: BACK-DOOR AND FRONT-DOOR CRITERIA

population of bird and other predators. This, in turn, is correlated with last year’s eelworm population and hence with the treatment itself.

The situation may be represented by the causal diagram in Figure 4.11. The variables are:

- $X$ fumigants,
- $Y$ crop yields,
- $Z_0$ last year’s eelworm population,
- $Z_1$ eelworm population before treatment,
- $Z_2$ eelworm population after treatment,
- $Z_3$ eelworm population at the end of the season,
- $B$ population of birds and other predators.

![Figure 4.11: A causal diagram representing the effect of fumigants $X$ on yields $Y$](image)

In this example, the variables $B$ and $Z_0$ are hidden variables.

The issue is whether interventional probabilities $\mathbb{P}_{Y \mid X}(\cdot \mid X=x)$ may be computed from information on the observables ($Z_1, Z_2, Z_3, X, Y$). When they can, they are said to be identifiable.

**Definition 4.4** (Identifiable). *The causal effect of $X$ on $Y$ is said to be identifiable if the quantity $\mathbb{P}_{Y \mid X}$ can be computed uniquely from the probability distribution of the observable variables.*

In this section, two graphical conditions are described which ensure that causal effects can be estimated consistently from observational data. The first of these is named back door criterion and is equivalent to the ignorance condition of Rosenbaum and Rubin. The second of these is the front-door criterion. This involves covariates which are affected by the treatment (in this example $Z_2$ and $Z_3$).
4.6.1 Back Door Criterion

The back door criterion is defined as follows:

**Definition 4.5** (Back Door Criterion). A set of nodes $C$ satisfies the back door criterion relative to an ordered pair of nodes $(X, Y) \in V \times V$ if

1. no node of $C$ is a descendant of $X$ and
2. $C$ blocks every trail (in the sense of D-separation) between $X$ and $Y$ which contains an edge pointing to $X$.

If $A$ and $B$ are two disjoint subsets of nodes, $C$ is said to satisfy the back door criterion relative to $(A, B)$ if it satisfies the back door criterion relative to any pair $(X_i, X_j) \in A \times B$.

**Example 4.4.**

In Figure 4.11, the set $C = \{Z_0\}$ satisfies the back door criterion relative to $(X, Y)$. The node $Z_0$ is unobservable. The set $C = \{Z_1, Z_2, Z_3\}$ does block all trails between $X$ and $Y$ with an arrow pointing into $X$ and hence satisfies the back door criterion.

The name ‘back door criterion’ reflects the fact that the second condition requires that only trails with nodes pointing at $X_i$ be blocked. The remaining trails can be seen as entering $X_i$ through a back door.

**Example 4.5.**

Consider the back door criterion DAG, given in Figure 4.12. The sets of variables $C_1 = \{Z_3, Z_4\}$ and $C_2 = \{Z_4, Z_3\}$ satisfy the back door criterion relative to the ordered pair of nodes $(X, Y)$, whereas $C_3 = \{Z_4\}$ does not satisfy the criterion relative to the ordered pair of nodes $(X, Y)$; if $Z_4$ is instantiated, the Bayes ball may pass through the collider connection from $Z_1$ to $Z_2$.

![Figure 4.12: Back Door Criterion](image-url)
Identifiability. Consider a causal network and $A$ a subset of the variables which satisfies the back door criterion with respect to an ordered pair $(X, Y)$. Such a set of variables $A$ plays a similar role to the variable $C$ in the discussion on confounding; if we can observe these variables, then we can estimate the intervention probability without a controlled experiment; otherwise we cannot.

If a set of variables $A$ satisfying the back door criterion with respect to $(X, Y)$ can be chosen such that $P_A$ and $P_{Y|A,X}$ can be estimated from the observed data, then the distribution $P_{Y|X}$ can also be estimated from the observed data.

Identifiability. If a set of variables $Z$ satisfies the back door criterion relative to $(X, Y)$, then the causal effect of $X$ to $Y$ is given by the formula

$$P_{Y|X} = (P_{Y|X,Z}P_Z)^{(X,Y)}$$

and the intervention of $X$ on $Y$ is said to be identifiable.

Formula (4.6) is named adjustment for concomitants. The word identifiability refers to the fact that the concomitants $Z$ satisfying the back door criterion are observable and hence it is possible to compute, or identify the intervention probability $P_{Y|X}(y|x)$ using the ‘see’ conditional probabilities $(P_{X_j|P_{A_j}})_{j=1}$.

4.6.2 Front Door Criterion

The front door criterion is defined as follows:

Definition 4.6 (Front Door Criterion). A set of variables $Z$ satisfies the front door criterion relative to the ordered pair $(X, Y)$ if:

- $Z$ intercepts all directed paths from $X$ to $Y$,
- there is no back-door path between $X$ and $Z$,
- every back-door path between $Z$ and $Y$ is blocked by $X$.

The situation is illustrated in Figure 4.13. The variable $U$ is a hidden (latent) variable. The variable $Z$ satisfies the front door criterion relative to $(X, Y)$.

![Figure 4.13: Front Door Criterion](image)

The result is the following:
\textbf{Theorem 4.7} (Front Door Criterion). \textit{Let }Z\textit{ satisfy the front door criterion relative to the ordered pair } (X, Y). \textit{Then the causal effect on } Y \textit{ of an intervention on } X \textit{ is:}

\[ P_{Y|X} = (P_{Z|X}P_{Y|Z})^{Z}. \]

This is self evident; note that \( P_{Y|Z} = (P_{Y|Z,U}P_{U})^{U(Y,Z)}. \) In other words, if the see-conditional \( P_{Z|X} \) and \( P_{Y|Z} \) are available, then the intervention \( P_{Y|X} \) may be computed. \( \square \)

\subsection{4.6.3 Non-Identifiability}

There are various conditions for non-identifiability of \( P_{Y|X}. \) These include:

1. A \textit{necessary} condition is that there is an unblockable back-door path between \( X \) and \( Y; \) that is, a path ending with an arrow pointing into \( X \) which cannot be blocked by observable non-descendants of \( X. \) This is not a sufficient condition, as Figure 4.13 illustrates. This shows a situation where there is a non-blockable back-door path, yet \( P_{Y|X} \) is identifiable (front-door criterion).

2. A \textit{sufficient} condition for identifiability of \( P_{Y|X} \) is existence of a confounding path between \( X \) and any of its children on a path from \( X \) to \( Y; \) two examples are given in Figure 4.14.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure414.png}
\caption{Sufficient condition for identifiability}
\end{figure}

3. Local identifiability is not a sufficient condition for global identifiability. In Figure 4.15, \( P_{Z_{1}|X}, P_{Y|Z_{1}}, P_{Y|Z_{2}} \) are all identifiable, but \( P_{Y|X} \) is not.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure415.png}
\caption{Sufficient condition for identifiability}
\end{figure}