CHAPTER 1

Causal Inference from Graphical Models

Steffen L. Lauritzen
Aalborg University

February 24, 2000

1.1 Introduction

The introduction of Bayesian networks (Pearl 1986b) and associated local computation algorithms (Lauritzen and Spiegelhalter 1988, Shenoy and Shafer 1990, Jensen, Lauritzen and Olesen 1990) has initiated a renewed interest for understanding causal concepts in connection with modelling complex stochastic systems.

It has become clear that graphical models, in particular those based upon directed acyclic graphs, have natural causal interpretations and thus form a base for a language in which causal concepts can be discussed and analysed in precise terms.

As a consequence there has been an explosion of writings, not primarily within mainstream statistical literature, concerned with the exploitation of this language to clarify and extend causal concepts. Among these we mention in particular books by Spirtes, Glymour and Scheines (1993), Shafer (1996), and Pearl (2000) as well as the collection of papers in Glymour and Cooper (1999).

Very briefly, but fundamentally, the important distinction to be made is the distinction between two types of conditional probability. We refer to these as conditioning by *intervention* and conditioning by *observation* and suggest the notation

$$p(x \,||\, y) = P(X = x \,|\, Y \leftarrow y), \quad p(x \,|\, y) = P(X = x \,|\, Y = y)$$

for these two notions. Other authors have used expressions such as $\mathrm{do}(y)$, $Y = \hat{y}$, and $\mathrm{set}(Y = y)$ to denote intervention conditioning.

Much existing controversy and lack of clarity is due to the misconception that these two are identical or even related in a simple fashion although the distinction has also been made both properly, clearly, and explicitly in better expositions of regression, see for example Box (1966) or Section 3.3 of Cox (1984).

In the following, we try to develop the basic ideas needed to make this distinction precise and discuss a number of classical statistical problems where the distinction is important.

There are many important aspects and views of causality and causal inference which are not even touched upon here, as we are only concerned with one particular such aspect: the prediction of the effect of interventions in a given system.

The material is organized as follows. Section 1.2 introduces the necessary graph-terminology. The next three sections are concerned with the very basic elements of graphical models, conditional independence and Markov properties for undirected and directed graphs.

Section 1.6 introduces the notion of a causal Markov field and associated intervention probabilities. The next sections are concerned with the exploitation of this idea in a number of important cases.

We conclude by discussing structural equation models and methods based upon using counterfactual variables or potential responses, and finally give a brief discussion of other issues which are not treated *per se* here.

While writing, I have in particular exploited Pearl (1995a) and Robins (1997).

1.2 Graph terminology

This section introduces some necessary graph terminology. We are basically following the terminology used in Cowell, Dawid, Lauritzen and Spiegelhalter (1999) which is almost identical to that in Lauritzen (1996).

We define a *graph* $\mathcal{G}$ to be a pair $\mathcal{G} = (V, E)$, where V is a finite set of *vertices*, also called *nodes*, of $\mathcal{G}$, and E is a subset of the set $V \times V$ of ordered pairs of vertices, called the *edges* or *links* of $\mathcal{G}$. Thus, as E is a set, the graph $\mathcal{G}$ has no multiple edges. We further require that E consist of pairs of distinct vertices, so that there are no loops.

If both ordered pairs (α, β) and (β, α) belong to E, we say that we have an *undirected* edge between α and β, and write $\alpha \sim \beta$; we also say that α and β are *neighbours*, α is a neighbour of β, or β is a neighbour of α. The set of neighbours of a vertex β is denoted by $\mathrm{ne}(\beta)$.

If $(\alpha, \beta) \in E$ but $(\beta, \alpha) \notin E$ we call the edge *directed*, and write $\alpha \rightarrow \beta$. We also say that α is a *parent* of β, and that β is a *child* of α. The set of parents of a vertex β is denoted by $\mathrm{pa}(\beta)$, and the set of children of a vertex α by $\mathrm{ch}(\alpha)$. The *family* of β, denoted $\mathrm{fa}(\beta)$, is $\mathrm{fa}(\beta) = \{\beta\} \cup \mathrm{pa}(\beta)$.

If $(\alpha,\beta) \in E$ or $(\beta,\alpha) \in E$ we say that α and β are *joined.* Then $\alpha \not\sim \beta$ indicates that α and β are not joined, i.e. both $(\alpha,\beta) \notin E$ and $(\beta,\alpha) \notin E$. We also write $\alpha \not\to \beta$ if $(\alpha,\beta) \notin E$.

If $A \subset V$, the expressions $\mathrm{pa}(A)$, $\mathrm{ne}(A)$ and $\mathrm{ch}(A)$ will denote the collection of parents, children and neighbours, respectively, of the elements of A, but excluding any element in A:

$$\begin{aligned} \mathrm{pa}(A) \quad &= \textstyle\bigcup_{\alpha \in A} \mathrm{pa}(\alpha) \setminus A, \\ \mathrm{ne}(A) \quad &= \textstyle\bigcup_{\alpha \in A} \mathrm{ne}(\alpha) \setminus A, \\ \mathrm{ch}(A) \quad &= \textstyle\bigcup_{\alpha \in A} \mathrm{ch}(\alpha) \setminus A. \end{aligned}$$

If all the edges of a graph are directed, we say that it is a *directed graph.* Conversely, if all the edges of a graph are undirected, we say that it is an *undirected graph.*

The *boundary* $\mathrm{bd}(\alpha)$ of a vertex α is the set of parents and neighbours of α; the boundary $\mathrm{bd}(A)$ of a subset $A \subset V$ is the set of vertices in $V \setminus A$ that are parents or neighbours to vertices in A, i.e. $\mathrm{bd}(A) = \mathrm{pa}\, A \cup \mathrm{ne}\, A$. The *closure* of A is $\mathrm{cl}(A) = A \cup \mathrm{bd}(A)$.

The *undirected version* $\mathcal{G}^{\sim}$ of a graph $\mathcal{G}$ is the undirected graph obtained by replacing the directed edges of $\mathcal{G}$ by undirected edges.

We call $\mathcal{G}_A = (A, E_A)$ a *subgraph* of $\mathcal{G} = (V, E)$ if $A \subseteq V$ and $E_A \subseteq E \cap (A \times A)$. Thus it may contain the same vertex set but possibly fewer edges. If in addition $E_A = E \cap (A \times A)$, we say that $\mathcal{G}_A$ is the subgraph of $\mathcal{G}$ *induced* by the vertex set A.

A graph is called *complete* if every pair of vertices are joined. We say that a subset of vertices of $\mathcal{G}$ is *complete* if it induces a complete subgraph. A complete subgraph which is maximal (with respect to $\subseteq$) is called a *clique.*

A *path* of length n from α to β is a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $(\alpha_{i-1}, \alpha_i) \in E$ for all $i = 1, \ldots, n$. Thus a path can never cross itself and moving along a path never goes against the directions of arrows.

A *cycle* of length n is a path with the modification that the first and last vertex are identical $\alpha_0 = \alpha_n$. The cycle is *directed* if it contains at least one arrow.

A directed graph which contains no cycles is called a *directed acyclic graph*, or DAG.

A *trail* of length n from α to β is a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $\alpha_{i-1} \to \alpha_i$, or $\alpha_i \to \alpha_{i-1}$, or $\alpha_{i-1} \sim \alpha_i$ for all $i = 1, \ldots, n$. Thus, moving along a trail could go against the direction of the arrows, in contrast to the case of a path. In other words, a trail in $\mathcal{G}$ is a sequence of vertices that form a path in the undirected version $\mathcal{G}^{\sim}$ of $\mathcal{G}$.

It is always possible to *well order* the nodes of a DAG, by a linear ordering or numbering, such that if two nodes are connected the edge points from the lower to the higher of the two nodes with respect to the ordering.

Given a directed acyclic graph, the set of its vertices α such that $\alpha \mapsto \beta$ but not $\beta \mapsto \alpha$ are the *ancestors* $\mathrm{an}(\beta)$ of β and the *descendants* $\mathrm{de}(\alpha)$ of α are the vertices β such that $\alpha \mapsto \beta$ but not $\beta \mapsto \alpha$. The *nondescendants* $\mathrm{nd}(\alpha)$ of α is the set $V \setminus (\mathrm{de}(\alpha) \cup \alpha)$. If $\mathrm{pa}(\alpha) \subseteq A$ for all $\alpha \in A$ we say that A is an *ancestral* set. The symbol $\mathrm{An}(A)$ denotes the smallest ancestral set containing A.

A subset $C \subseteq V$ is said to be an (α, β)-*separator* if all trails from α to β intersect C. The subset C is said to *separate* A from B if it is an (α, β)-separator for every $\alpha \in A$ and $\beta \in B$. An (α, β)-separator C is said to be *minimal* if no proper subset of C is itself an (α, β)-separator.

For a directed acyclic graph $\mathcal{D}$, we define the *moral graph* of $\mathcal{D}$ to be the undirected graph $\mathcal{D}^m$ obtained from $\mathcal{D}$ by first adding undirected edges between all pairs of vertices which have common children and are not already joined, and then forming the undirected version of the resulting graph.

1.3 Conditional independence

Throughout this text a central notion is that of conditional independence of random variables and groups of these.

We are concerned with the situation where we have a collection of random variables $(X_\alpha)_{\alpha \in V}$ taking values in probability spaces $(\mathcal{X}_\alpha)_{\alpha \in V}$. The probability spaces are either real finite-dimensional vector spaces or finite and discrete sets but could be quite general, just sufficiently well-behaved to ensure the existence of conditional probabilities. For simplicity we mostly consider the discrete case.

For A being a subset of V we let $\mathcal{X}_A = \times_{\alpha \in A} \mathcal{X}_\alpha$ and further $\mathcal{X} = \mathcal{X}_V$. Typical elements of $\mathcal{X}_A$ are denoted as $x_A = (x_\alpha)_{\alpha \in A}$. Similarly $X_A = (X_\alpha)_{\alpha \in A}$.

Formally, if X, Y, Z are random variables with a joint distribution P, we say that X *is conditionally independent of* Y *given* Z *under* P, and write $X \perp\!\!\!\perp Y \mid Z\ [P]$, if, for any measurable set A in the sample space of X, there exists a version of the conditional probability $P(A \mid Y, Z)$ which is a function of Z alone. Usually P will be fixed and omitted from the notation. If Z is trivial we say that X *is independent of* Y, and write $X \perp\!\!\!\perp Y$.

When X, Y, and Z are discrete random variables the condition for $X \perp\!\!\!\perp Y \mid Z$ simplifies as

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z) P(Y = y \mid Z = z),$$

where the equation holds for all z with $P(Z = z) > 0$. When the three variables admit a joint density with respect to a product measure μ, we have

$$X \perp\!\!\!\perp Y \mid Z \iff f_{XY \mid Z}(x, y \mid z) = f_{X \mid Z}(x \mid z) f_{Y \mid Z}(y \mid z), \qquad (1.1)$$

where this equation is to hold almost surely with respect to P. The condition (1.1) can be rewritten as

$$X \perp\!\!\!\perp Y \mid Z \iff f_{XYZ}(x,y,z)f_Z(z) = f_{XZ}(x,z)f_{YZ}(y,z) \tag{1.2}$$

and this equality must hold *for all values of* z when the densities are continuous.

The ternary relation $X \perp\!\!\!\perp Y \mid Z$ has the following properties, where h denotes an arbitrary measurable function on the sample space of X:

(C1) if $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$;

(C2) if $X \perp\!\!\!\perp Y \mid Z$ and $U = h(X)$, then $U \perp\!\!\!\perp Y \mid Z$;

(C3) if $X \perp\!\!\!\perp Y \mid Z$ and $U = h(X)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$;

(C4) if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (W, Y) \mid Z$.

Note that the converse to (C4) follows from (C2) and (C3).

If we use f as generic symbol for the probability density of the random variables corresponding to its arguments, the following statements are true:

$$\begin{aligned}
X \perp\!\!\!\perp Y \mid Z &\iff f(x,y,z) = f(x,z)f(y,z)/f(z) && (1.3)\\
X \perp\!\!\!\perp Y \mid Z &\iff f(x \mid y,z) = f(x \mid z) && (1.4)\\
X \perp\!\!\!\perp Y \mid Z &\iff f(x,z \mid y) = f(x \mid z)f(z \mid y) && (1.5)\\
X \perp\!\!\!\perp Y \mid Z &\iff f(x,y,z) = h(x,z)k(y,z) \text{ for some } h,k && (1.6)\\
X \perp\!\!\!\perp Y \mid Z &\iff f(x,y,z) = f(x \mid z)f(y,z). && (1.7)
\end{aligned}$$

The equalities above hold apart from a set of triples (x,y,z) with probability zero.

Another property of the conditional independence relation is often used:

(C5) if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$ then $X \perp\!\!\!\perp (Y, Z)$.

However (C5) does not hold universally, but only under additional conditions — essentially that there be no non-trivial logical relationship between Y and Z. A trivial counterexample appears when $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = 1/2$. We have however

Proposition 1.1 *If the joint density of all variables with respect to a product measure is strictly positive, then the statement* (C5) *will hold true.*

Proof: We assume for simplicity that the variables are discrete with density $f(x,y,z) > 0$ and that $X \perp\!\!\!\perp Y \mid Z$ as well as $X \perp\!\!\!\perp Z \mid Y$. Then (1.6) gives for all values of (x,y,z) that

$$f(x,y,z) = k(x,z)l(y,z) = g(x,y)h(y,z)$$

for suitable strictly positive functions g, h, k, l. Thus we have for all z that

$$g(x,y) = \frac{k(x,z)l(y,z)}{h(y,z)}.$$

Choosing a fixed $z = z_0$ we get $g(x, y) = \pi(x)\rho(y)$ where $\pi(x) = k(x, z_0)$ and $\rho(y) = l(y, z_0)/h(y, z_0)$. Thus $f(x, y, z) = \pi(x)\rho(y)h(y, z)$ and hence $X \perp\!\!\!\perp (Y, Z)$ as desired. □

In most cases we are specifically interested in conditional independence among groups of random variables such as for example $X_A = (X_\alpha, \alpha \in A)$, where A is a subset of V. We then use the short notation

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

for

$$X_A \perp\!\!\!\perp X_B \mid X_C$$

and so on. We then get the following properties as a consequence of (C1)–(C4):

(C1') if $A \perp\!\!\!\perp B \,|\, C$ then $B \perp\!\!\!\perp A \,|\, C$;

(C2') if $A \perp\!\!\!\perp B \,|\, C$ and $D \subseteq B$, then $A \perp\!\!\!\perp D \,|\, C$;

(C3') if $A \perp\!\!\!\perp B \,|\, C$ and $D \subseteq B$, then $A \perp\!\!\!\perp B \,|\, C \cup D$;

(C4') if $A \perp\!\!\!\perp B \,|\, C$ and $A \perp\!\!\!\perp D \,|\, B \cup C$, then $A \perp\!\!\!\perp B \cup D \,|\, C$.

And similarly the analogue of (C5) is that for disjoint subsets A, B, C, and D, we have

(C5') if $A \perp\!\!\!\perp B \,|\, C \cup D$ and $A \perp\!\!\!\perp C \,|\, B \cup D$ then $A \perp\!\!\!\perp B \cup C \,|\, D$

although (C5') does not hold universally, but only under specific extra assumptions. It holds for example under the assumption that the joint density of the random variables involved is strictly positive.

It is illuminating to think of the properties (C1)–(C5) or in particular their analogues (C1')–(C5') as purely formal expressions, with a meaning that is not necessarily tied to probability. If we interpret the symbols used for random variables as abstract symbols for pieces of knowledge obtained from, say, reading books, and further interpret the symbolic expression $X \perp\!\!\!\perp Y \,|\, Z$ as:

Knowing Z, reading Y is irrelevant for reading X,

the properties (C1)–(C4) translate to the following:

(I1) if, knowing Z, reading Y is irrelevant for reading X, then so is reading X for reading Y;

(I2) if, knowing Z, reading Y is irrelevant for reading the book X, then reading Y is irrelevant for reading any chapter U of X;

(I3) if, knowing Z, reading Y is irrelevant for reading the book X, it remains irrelevant after having read any chapter U of X;

(I4) if, knowing Z, reading the book Y is irrelevant for reading X and even after having also read Y, reading W is irrelevant for reading X, then reading of both Y and W is irrelevant for reading X.

Thus one can view the relations (C1)–(C4) as pure formal properties of the notion of irrelevance. The property (C5) is slightly more subtle. In a certain sense, also the symmetry (C1) is a somewhat special property of probabilistic conditional independence, rather than general irrelevance.

It is thus tempting to use the relations (C1)–(C4) as formal axioms for conditional independence or irrelevance. A *semi-graphoid* is an algebraic structure which satisfies (C1')–(C4'). If also (C5') holds for disjoint subsets, it is called a *graphoid* (Pearl 1988). Similarly we refer to (C1')–(C4') as the *semi-graphoid axioms* and (C1')–(C5') as the *graphoid axioms.*

1.4 Markov properties for undirected graphs

Conditional independence properties of joint distributions of collections of random variables can be compactly described and expressed as so-called Markov properties for various graphs. In this section we consider the case when the graph is undirected. We refer to Lauritzen (1996) or Cowell et al. (1999) for proofs of all assertions that are not proved here.

Associated with an undirected graph $\mathcal{G} = (V, E)$ and a collection of random variables $(X_\alpha)_{\alpha \in V}$ as above there is a range of different Markov properties. A probability distribution P on $\mathcal{X}$ is said to obey

(P) *the pairwise Markov property*, relative to $\mathcal{G}$, if for any pair (α, β) of non-adjacent vertices

$$\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\};$$

(L) *the local Markov property*, relative to $\mathcal{G}$, if for any vertex $\alpha \in V$

$$\alpha \perp\!\!\!\perp V \setminus \mathrm{cl}(\alpha) \mid \mathrm{bd}(\alpha);$$

(G) *the global Markov property*, relative to $\mathcal{G}$, if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $\mathcal{G}$

$$A \perp\!\!\!\perp B \mid S.$$

As conditional independence is intimately related to factorization, so are the Markov properties. A probability measure P on $\mathcal{X}$ is said to *factorize* according to $\mathcal{G}$ if for all complete subsets $a \subseteq V$ there exist non-negative functions ψ_a that depend on x through x_a only, and there exists a product measure $\mu = \otimes_{\alpha \in V} \mu_\alpha$ on $\mathcal{X}$, such that P has density f with respect to μ where f has the form

$$f(x) = \prod_{a \text{ complete}} \psi_a(x). \tag{1.8}$$

The functions ψ_a are not uniquely determined. There is arbitrariness in the choice of μ, but also groups of functions ψ_a can be multiplied together or split up in different ways. In fact one can without loss of generality assume

— although this is not always practical — that only cliques appear as the sets a, i.e. that

$$f(x) = \prod_{c \in \mathcal{C}} \psi_c(x), \tag{1.9}$$

where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$. If P factorizes, we say that P has property (F). The different Markov properties are related as follows

Proposition 1.2 *For any undirected graph $\mathcal{G}$ and any probability distribution on $\mathcal{X}$ it holds that*

$$(\mathrm{F}) \implies (\mathrm{G}) \implies (\mathrm{L}) \implies (\mathrm{P}).$$

Proof: See Lauritzen (1996). □

For a given graph $\mathcal{G}$ and state space $\mathcal{X} = \times_{\alpha \in V} \mathcal{X}_\alpha$ we denote the set of distributions that satisfy the different Markov properties as $M_F(\mathcal{G})$, $M_G(\mathcal{G})$, $M_L(\mathcal{G})$, and $M_P(\mathcal{G})$. Proposition 1.2 can now be equivalently formulated as

$$M_F(\mathcal{G}) \subseteq M_G(\mathcal{G}) \subseteq M_L(\mathcal{G}) \subseteq M_P(\mathcal{G}).$$

The Markov properties are genuinely different in general, but in the case where P has a positive density it is possible to show that (P) implies (F), and thus that all Markov properties are equivalent. This result has been discovered in various forms by a number of authors (Speed 1979) but is usually attributed to Hammersley and Clifford (1971). More precisely, we have

Theorem 1.3 (Hammersley and Clifford) *A probability distribution P with positive density f with respect to a product measure μ satisfies the pairwise Markov property with respect to an undirected graph $\mathcal{G}$ if and only if it factorizes according to $\mathcal{G}$.*

Proof: See Lauritzen (1996). □

In fact, if (C5') holds, the global, local, and pairwise Markov properties coincide. This fact is stated in the theorem below, due to Pearl and Paz (1987); see also Pearl (1988).

Theorem 1.4 (Pearl and Paz) *If a probability distribution on $\mathcal{X}$ is such that (C5') holds for disjoint subsets A, B, C, D then*

$$(\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{P}).$$

Proof: See Lauritzen (1996). □

The global Markov property (G) is important because it gives a general criterion for deciding when two groups of variables A and B are conditionally independent given a third group of variables S. Moreover, it cannot be further strengthened. For example it holds (Frydenberg 1990b) that if all state spaces are binary, i.e. $\mathcal{X}_\alpha = \{1, -1\}$, then

$$A \perp\!\!\!\perp B \mid S \text{ for all } P \in M_F(\mathcal{G}) \iff S \text{ separates } A \text{ from } B.$$

In other words, if A and B are not separated by S then there is a factorizing distribution that makes them conditionally dependent.

1.5 The directed Markov property

We consider the same set-up as in the previous section, except that now the graph $\mathcal{D}$ is assumed to be directed and acyclic.

We say that a probability distribution P admits a *recursive factorization* according to $\mathcal{D}$, if there exist (σ-finite) measures μ_α over $\mathcal{X}$ and non-negative functions $k^\alpha(\cdot,\cdot), \alpha \in V$, henceforth referred to as *kernels*, defined on $\mathcal{X}_\alpha \times \mathcal{X}_{\mathrm{pa}(\alpha)}$ such that

$$\int k^\alpha(y_\alpha, x_{\mathrm{pa}(\alpha)})\mu_\alpha(dy_\alpha) = 1$$

and P has density f with respect to the product measure $\mu = \otimes_{\alpha\in V}\mu_\alpha$ given by

$$f(x) = \prod_{\alpha\in V} k^\alpha(x_\alpha, x_{\mathrm{pa}(\alpha)}).$$

We then also say that P *has property* (DF). It is easy to show that, if P admits a recursive factorization as above, then the kernels $k^\alpha(\cdot, x_{\mathrm{pa}(\alpha)})$ are in fact densities for the conditional distribution of X_α, given $X_{\mathrm{pa}(\alpha)} = x_{\mathrm{pa}(\alpha)}$ and thus

$$f(x) = \prod_{\alpha\in V} f(x_\alpha \mid x_{\mathrm{pa}(\alpha)}). \tag{1.10}$$

We refer to these kernels as the *conditional specifications* for P. It is immediate that if we form the (undirected) moral graph $\mathcal{D}^m$ (see Section 1.2) we have the following:

Lemma 1.5 *If P admits a recursive factorization according to the directed acyclic graph $\mathcal{D}$, it factorizes according to the moral graph $\mathcal{D}^m$ and therefore obeys the global Markov property relative to $\mathcal{D}^m$.*

Proof: The factorization follows from the fact that, by construction, the sets $\{\alpha\}\cup\mathrm{pa}(\alpha)$ are complete in $\mathcal{D}^m$ and we can therefore let $\psi_{\{\alpha\}\cup\mathrm{pa}(\alpha)} = k^\alpha$. □

This simple lemma has very useful consequences and we shall see several examples of this in the sequel. Also, using the local Markov property on the moral graph $\mathcal{D}^m$ we find that

$$\alpha \perp\!\!\!\perp V\setminus\alpha \mid \mathrm{bl}(\alpha),$$

where $\mathrm{bl}(\alpha)$ is the so-called *Markov blanket* of α. The Markov blanket is the set of neighbours of α in the moral graph $\mathcal{D}^m$. It can be found directly from the original DAG $\mathcal{D}$ as the set of α's parents, children, and children's parents:

$$\mathrm{bl}(\alpha) = \mathrm{pa}(\alpha) \cup \mathrm{ch}(\alpha) \cup \{\beta : \mathrm{ch}(\beta) \cap \mathrm{ch}(\alpha) \neq \emptyset\}. \tag{1.11}$$

In particular it follows that the so-called full conditionals satisfy

$$\mathcal{L}(X_\alpha \,|\, X_{V\setminus\alpha}) = \mathcal{L}(X_\alpha \,|\, X_{\mathrm{bl}(\alpha)})$$

with density given as

$$\mathcal{L}(X_\alpha \,|\, X_{V\setminus\alpha}) = f(x_\alpha \,|\, x_{\mathrm{pa}(\alpha)}) \prod_{\beta\in\mathrm{ch}(\alpha)} f(x_\beta \,|\, x_{\mathrm{pa}(\beta)}).$$

The following result is easily shown:

Proposition 1.6 *If P admits a recursive factorization according to the directed acyclic graph $\mathcal{D}$ and A is an ancestral set, then the marginal distribution P_A admits a recursive factorization according to $\mathcal{D}_A$.*

In combination with Lemma 1.5 this yields:

Corollary 1.7 *Let P factorize recursively according to $\mathcal{D}$. Then*

$$A \perp\!\!\!\perp B \,|\, S$$

whenever A and B are separated by S in $(\mathcal{D}_{\mathrm{An}(A\cup B\cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$.

Following Lauritzen, Dawid, Larsen and Leimer (1990), the property in Corollary 1.7 will be referred to as the *directed global Markov property* (DG) and a distribution satisfying it is a *directed Markov field* over $\mathcal{D}$.

One can show that the global directed Markov property has the same rôle as the global Markov property does in the case of an undirected graph, in the sense that it gives the sharpest possible rule for reading conditional independence relations off the directed graph. The procedure is illustrated in the following example:

Example 1.8 Consider a directed Markov field on the first graph in Fig. 1.1 and the problem of deciding whether $a \perp\!\!\!\perp b \,|\, S$. The moral graph of the smallest ancestral set containing all the variables involved is shown in the second graph of Fig. 1.1. It is immediate that S separates a from b in this moral graph, implying $a \perp\!\!\!\perp b \,|\, S$. □

An alternative formulation of the global, directed Markov property was given by Pearl (1986a) with a formal treatment in Verma and Pearl (1990). Recall that a trail in $\mathcal{D}$ is a sequence of vertices that forms a path in the undirected version $\mathcal{D}^\sim$ of $\mathcal{D}$, i.e. when the directions of arrows are ignored. A trail π from a to b in a directed, acyclic graph $\mathcal{D}$ is said to be *blocked* by S if it contains a vertex $\gamma \in \pi$ such that either

$\gamma \in S$ and arrows of π do not meet head-to-head at γ, or

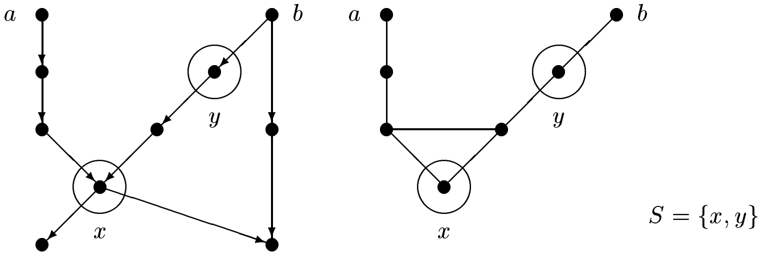


Figure 1.1 *The directed, global Markov property. Is $a \perp\!\!\!\perp b \mid S$? In the moral graph of the smallest ancestral set in the graph containing $\{a\} \cup \{b\} \cup S$, clearly S separates a from b, implying $a \perp\!\!\!\perp b \mid S$.*

γ and all its descendants are not in S, and arrows of π meet head-to-head at γ.

A trail that is not blocked by S is said to be *active*. Two subsets A and B are said to be *d-separated* by S if all trails from A to B are blocked by S. We then have the following result:

Proposition 1.9 *Let A, B and S be disjoint subsets of a directed, acyclic graph $\mathcal{D}$. Then S d-separates A from B if and only if S separates A from B in $(\mathcal{D}_{\mathrm{An}(A \cup B \cup S)})^m$.*

Proof: See Lauritzen (1996). □

The global directed Markov property can thus be formulated by requiring that $A \perp\!\!\!\perp B \mid S$ whenever S d-separates A from B thereby making the analogy with the undirected case clearer. It depends on the specific context and purpose whether the pathwise criterion, or the criterion used in the definition of the global directed Markov property is easiest to use.

We illustrate the concept of d-separation by applying it to the query of Example 1.8. As Fig. 1.2 indicates, all trails between a and b are blocked by S, whereby the global Markov property gives that $a \perp\!\!\!\perp b \mid S$.

For further use, we shall use the symbolic expression $A \perp_{\mathcal{D}} B \mid S$ to denote that A and B are d-separated by S or, equivalently, A and B are separated by S in $(\mathcal{D}_{\mathrm{An}(A \cup B \cup S)})^m$. It was shown in Verma and Pearl (1990) that

Lemma 1.10 *For any fixed directed acyclic graph $\mathcal{D}$, the relation $\perp_{\mathcal{D}}$ satisfies the graphoid axioms.*

Geiger and Pearl (1990) show that the criterion of d-separation cannot be improved, in the sense that, for any given directed acyclic graph $\mathcal{D}$, one

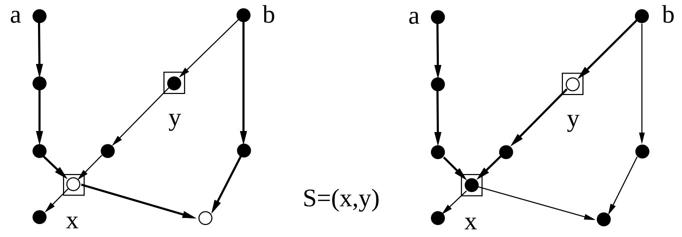


Figure 1.2 *Illustration of Pearl's d-separation criterion. There are two trails from a to b, drawn with thick lines. Both are blocked, but different vertices γ, indicated with open circles, play the rôle of blocking vertices.*

can find state spaces $\mathcal{X}_\alpha, \alpha \in V$ and a probability distribution P such that

$$A \perp\!\!\!\perp B \mid S \iff A \perp_{\mathcal{D}} B \mid S. \tag{1.12}$$

This result was strengthened by Meek (1995), who showed that if the state spaces were finite and had cardinality at least two, the set of probability distributions P not satisfying (1.12) had Lebesgue measure zero in the set of all directed Markov probability measures.

To complete this section we say that P obeys the *local directed Markov property* (DL) if any variable is conditionally independent of its non-descendants, given its parents:

$$\alpha \perp\!\!\!\perp \mathrm{nd}(\alpha) \mid \mathrm{pa}(\alpha).$$

A seemingly weaker requirement, the *ordered directed Markov property* (DO), replaces all non-descendants of α in the above condition by the predecessors $\mathrm{pr}(\alpha)$ of α in some given well-ordering of the nodes:

$$\alpha \perp\!\!\!\perp \mathrm{pr}(\alpha) \mid \mathrm{pa}(\alpha).$$

In contrast with the undirected case we have that all the four properties (DF), (DL), (DG) and (DO) are equivalent just assuming existence of the density f. This is stated formally as:

Theorem 1.11 *Let $\mathcal{D}$ be a directed acyclic graph. For a probability distribution P on $\mathcal{X}$ which has density with respect to a product measure μ, the following conditions are equivalent:*

(DF) *P admits a recursive factorization according to $\mathcal{D}$;*
(DG) *P obeys the global directed Markov property, relative to $\mathcal{D}$;*
(DL) *P obeys the local directed Markov property, relative to $\mathcal{D}$;*
(DO) *P obeys the ordered directed Markov property, relative to $\mathcal{D}$.*

Proof: That (DF) implies (DG) is Corollary 1.7. That (DG) implies (DL) follows by observing that $\{\alpha\} \cup \mathrm{nd}(\alpha)$ is an ancestral set and that $\mathrm{pa}(\alpha)$ obviously separates $\{\alpha\}$ from $\mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha)$ in $(\mathcal{D}_{\{\alpha\}\cup\mathrm{nd}(\alpha)})^m$. It is trivial that (DL) implies (DO), since $\mathrm{pr}(\alpha) \subseteq \mathrm{nd}(\alpha)$. The final implication is shown by induction on the number of vertices $|V|$ of $\mathcal{D}$. Let α_0 be the last vertex of $\mathcal{D}$. Then we can let k^{α_0} be the conditional density of X_{α_0}, given $X_{V\setminus\{\alpha_0\}}$, which by (DO) can be chosen to depend on $x_{\mathrm{pa}(\alpha_0)}$ only. The marginal distribution of $X_{V\setminus\{\alpha_0\}}$ trivially obeys the ordered directed Markov property and admits a factorization by the inductive assumption. Combining this factorization with k^{α_0} yields the factorization for P. This completes the proof. □

Since the four conditions in Theorem 1.11 are equivalent, it makes sense to speak of a *directed Markov field* as one where any of the conditions is satisfied. The set of such distributions for a directed acyclic graph $\mathcal{D}$ is denoted by $M(\mathcal{D})$.

In the particular case when the directed acyclic graph $\mathcal{D}$ is perfect, i.e. all parents are married, the directed Markov property on $\mathcal{D}$ and the factorization property on its undirected version $\mathcal{D}^{\sim}$ coincide.

Proposition 1.12 *Let $\mathcal{D}$ be a perfect directed acyclic graph and $\mathcal{D}^{\sim}$ its undirected version. Then P is directed Markov with respect to $\mathcal{D}$ if and only if it factorizes according to $\mathcal{D}^{\sim}$.*

Proof: See Lauritzen (1996). □

1.6 Causal Markov models

For simplicity we assume here and in the following that all random variables are discrete and have finite state spaces unless we specifically indicate otherwise. To emphasize the discreteness we use little p as a generic symbol for a probability mass function rather than f for a general density.

1.6.1 Conditioning by observation or intervention

The first important issue is to distinguish between different types of conditioning operations, each of which modify a given probability distribution in response to information obtained. Conditional probabilities are sometimes defined and calculated as

$$p(y \mid x) = P(Y = y \mid X = x) = \frac{P(Y = y, X = x)}{P(X = x)}.$$

We refer to this type of conditioning as *conditioning by observation* or *conventional conditioning*. In many cases this represents the way in which a probability distribution, $P(Y = y)$, should be modified when the information $X = x$ is revealed. Paradoxes appear when it is unclear how the information about X is revealed (Shafer 1985, 1996), but that is a different discussion.

When discussing causal issues it is important to realize that this is typically not the way the distribution of Y should be modified if we intervene externally and force the value of X to be equal to x. We refer to this type of modification as *conditioning by intervention* or *conditioning by action*. To make the distinction clear we use different symbols for this conditioning, as indicated below

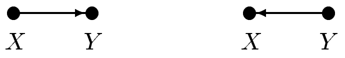
$$p(y \,||\, x) = P(Y = y \,|\, X \leftarrow x).$$

Generally, the two quantities will be different

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

and the quantity on the left-hand side cannot be calculated from the probability measure P alone, without additional assumptions. To judge whether these assumptions are reasonable in any given context one needs a specification of the precise way in which the intervention is made, just as conventional conditioning needs a specification about how the information is revealed.

In a moment we will give a precise meaning to a directed acyclic graph being causal. This will imply that in the graph below to the left



we will have that $p(y \,||\, x) = p(y \,|\, x)$ and $p(x \,||\, y) = p(x)$, whereas these relations are reversed in the graph to the right, i.e. there it holds that $p(y \,||\, x) = p(y)$ and $p(x \,||\, y) = p(x \,|\, y)$.

1.6.2 Causal graphs

A directed acyclic graph $\mathcal{D}$ is said to be *causal* for a probability distribution P with respect to a subset $B \subseteq V$, if P is Markov with respect to $\mathcal{D}$, i.e.

$$p(x) = \prod_{\alpha \in V} p(x_\alpha \,|\, x_{\mathrm{pa}(\alpha)})$$

and it further holds for any $A \subseteq B$ that

$$\begin{aligned} p(x \,||\, x_A^*) &= \prod_{\alpha \in V \setminus A} p(x_\alpha \,|\, x_{\mathrm{pa}(\alpha)}) \Bigg|_{x_A = x_A^*} \\ &= \frac{p(x)}{\prod_{\alpha \in A} p(x_\alpha^* \,|\, x_{\mathrm{pa}(\alpha)})} \Bigg|_{x_A = x_A^*} . \end{aligned} \qquad (1.13)$$

Alternatively, one can think of the right-hand side of (1.13) as the mathematical definition of the intervention probability on the left-hand side.

If $B = V$ we simply say that $\mathcal{D}$ is *causal* or *fully causal* for P. We also use the expression that P is a *causal directed Markov field* with respect to $\mathcal{D}$ or say that P is *causally Markov* with respect to $\mathcal{D}$. Note that the causal Markov property thus gives a relation between different probability measures, each representing the probability law associated with a specific intervention.

We will refer to (1.13) as the *intervention formula.* It appeared in various forms in Pearl (1993) and Spirtes et al. (1993). It is implicit in Robins (1986) and in other literature.

There are many ways in which this causal interpretation of a directed Markov model can be justified. But it is also important to realize that there are many other ways in which one can associate causal relationships with directed acyclic graphs. This is in particular apparent in the highly interesting book of Shafer (1996) who develops a language for causal interpretation of probabilities through event trees. This leads to events being more natural as direct causes than variables. A variety of causal relationships between variables can then be derived as consequences of the formalism.

In a more general setting one would be interested in allowing other types of intervention than those described. For example, one could wish to control the value of a variable in a way that depends on previously observed variables. But for simplicity we only consider the case of simple interventions.

One should contrast the intervention formula (1.13) with conventional conditioning using Bayes' formula:

$$p(x \mid x_A^*) = \left.\frac{p(x)}{p(x_A^*)}\right|_{x_A = x_A^*} = \left.\frac{p(x)}{\sum_{y : y_A = x_A^*} p(y)}\right|_{x_A = x_A^*}, \tag{1.14}$$

which differs from the intervention formula in the denominator, where the product of conditional specifications is replaced by the marginal probability $p(x_A^*)$. This implies in particular that if intervention takes place on a single variable without parents, observation and intervention have identical effects:

Corollary 1.13 *If* $\alpha \in V$ *has no parents, i.e.* $\mathrm{pa}(\alpha) = \emptyset$*, then it holds that* $p(x \,||\, x_\alpha^*) = p(x \mid x_\alpha^*)$.

We illustrate the similarities and differences by intervening on variable 5 in Figure 1.3. If this graph is causal, we have that the intervention $X_5 \leftarrow x_5^*$ produces the distribution

$$p(x \,||\, x_5^*) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_6 \mid x_3, x_5^*)p(x_7 \mid x_4, x_5^*, x_6)$$

whereas the observation $X_5 = x_5^*$ leads to

$$p(x \mid x_5^*) \quad \propto \quad p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)$$

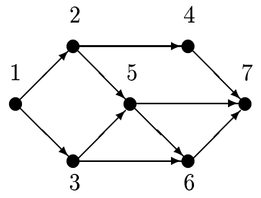


Figure 1.3 *Illustration of causal directed acyclic graph.*

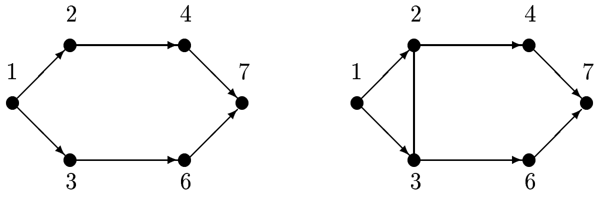


Figure 1.4 *The intervention $X_5 \leftarrow x_5^*$ in Fig. 1.3 produces a causal directed Markov field with respect to the graph on the left. The observation $X_5 = x_5^*$ produces a distribution which satisfies the chain graph Markov property with respect to the graph to the right.*

$$p(x_5^* \mid x_2, x_3) p(x_6 \mid x_3, x_5^*) p(x_7 \mid x_4, x_5^*, x_6). \qquad (1.15)$$

The modified distribution $P(\cdot \mid X_A \leftarrow x_A^*)$ is again a directed causal Markov field over the subgraph $\mathcal{D}_{V \setminus A}$ induced by the remaining variables. The corresponding conditional specifications are just modified such that

$$p(x_\alpha \mid x_{\mathrm{pa}_{V \setminus A}(\alpha)} \,||\, x_A^*) = p(x_\alpha \mid x_{\mathrm{pa}(\alpha) \setminus A}, x^*_{\mathrm{pa}(\alpha) \cap A}).$$

Expressed in words, the causal assumption is that the conditional specifications are unchanged for variables which are not used for intervention. In the example above, where we have intervened on variable 5, the only modifications of the specifications involve children of the intervention variables, i.e. variables 6 and 7, where we get

$$p(x_6 \mid x_3 \,||\, x_5^*) = p(x_6 \mid x_3, x_5^*), \quad p(x_7 \mid x_4, x_6 \,||\, x_5^*) = p(x_7 \mid x_4, x_5^*, x_6).$$

The corresponding subgraph is displayed as the graph to the left in Fig. 1.4. This is again to be contrasted with the effect of observation of variable 5, which creates a dependence structure determined by the chain graph (Lauritzen 1996) to the right in the same figure. This is due to the factor $p(x_5^* \mid x_2, x_3)$ creating a function depending on (x_2, x_3) in the factorization (1.15). Note that, in general, the conditional independence structure induced by conditioning by observation will not be in perfect correspondence with a chain graph.

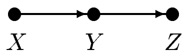
It is important to realize that successive conditioning operations of the same type commute whereas intervention and observation in general cannot be interchanged. We therefore adopt the convention that all operations are performed from the right to the left. Thus

$$p(x \,||\, y \,|\, z) = P(X = x \,|\, Y \leftarrow y, Z = z)$$

denotes the modified probability obtained by first observing $Z = z$ and subsequently intervening as $Y \leftarrow y$, whereas

$$p(x \,|\, z \,||\, y) = P(X = x \,|\, Z = z, Y \leftarrow y)$$

reflects that the intervention is performed before the observation. As an example, consider the graph



Intervening with $X \leftarrow x^*$ and then subsequently observing $Y = y$ leads to

$$p(z \,|\, y \,||\, x^*) = p(z \,|\, y),$$

whereas additional assumptions are to be made to predict the effect of the intervention $X \leftarrow x^*$ after observation of $Y = y$. Such assumptions could for example be that X, Y, and Z are functionally related in a structural equation model, see Section 1.9. This assumption would lead to the equality

$$p(z \,||\, x^* \,|\, y) = p(z \,|\, y \,||\, x^*) = p(z \,|\, y)$$

as then X and Z are functionally unrelated, once the value of y is known or has been fixed.

Generally, to ensure unambiguous meaning of intervention conditioning without introducing assumptions beyond those already made, intervention at a node α must always be made before any variables corresponding to its descendant nodes have been observed.

1.7 Assessment of treatment effects in sequential trials

The following example is adapted from Robins (1997) and is the simplest example where traditional approaches to assessment of treatment effects give incorrect results, whereas the methods described here coincide with those developed by Robins (1986), known as G-computation, and give the correct answer.

Consider a study made in a population of AIDS patients. Let us imagine the population being so large that sampling error can be ignored for practical purposes. The study involves 4 binary variables. In our notation, a is the label for an initial, randomized treatment, where $X_a = 1$ denotes that the patient has been treated with AZT, and $X_a = 0$ indicates placebo. After a given period it is for each patient observed whether

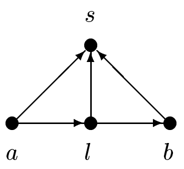


Figure 1.5 *Graph displaying causal relationships between variables in a particular sequential trial. The graph is only assumed causal with respect to interventions at* $\{a, b\}$. *The missing arrow from* a *to* b *reflects that* b *is assigned by randomization.*

the patient develops pneumonia, corresponding to the variable l, where $X_l = 1$ indicates that this is the case. We assume that all patients survive up to this point. Subsequently a secondary treatment with antibiotics is contemplated, corresponding to the variable b. For ethical reasons, all patients who have developed pneumonia are treated with antibiotics, i.e. $P(X_b = 1 \mid X_l = 1) = 1$, whereas the treatment is randomized for the patients with $X_l = 0$. Finally, after a given period it is registered whether a given patient has survived up to that time, corresponding to the variable s, where $X_s = 1$ denotes that the patient has survived.

The question is now to assess the effect on survival of a combined treatment with AZT and antibiotics of a new patient. In other words, we wish to calculate

$$P(X_s = 1 \mid X_a \leftarrow 1, X_b \leftarrow 1) = p(1_s \,||\, 1_a, 1_b).$$

This is done in the following way. The relevant graph is displayed in Figure 1.5, where missing arrows reflect the randomized allocation of treatments. This graph is not fully causal as there may be unobserved variables (confounders) that simultaneously affect l and r. It is only assumed causal with respect to interventions at $\{a, b\}$.

Note that not all effects are estimable as there are no observations with $X_l = 1$ and $X_b = 0$. For example, the effect of treating with AZT only cannot be assessed. We find

$$\begin{aligned} p(1_s \,||\, 1_a, 1_b) &= \sum_{x_l} p(1_s, x_l \,||\, 1_a, 1_b) \\ &= \sum_{x_l} p(1_s \mid x_l \,||\, 1_a, 1_b) p(x_l \,||\, 1_a, 1_b) \\ &= \sum_{x_l} p(1_s \mid x_l, 1_a, 1_b) p(x_l \mid 1_a). \end{aligned}$$

As pointed out by Robins, conventional wisdom gives ambiguous or incorrect answers: The variable l is affected by the treatment and one should therefore not adjust for it but simply use the estimate of the conditional probability

$$p(1_s \,||\, 1_a, 1_b) \sim \hat{p}(1_s \,|\, 1_a, 1_b).$$

On the other hand, the covariate l is also a confounder for the effect of the treatment on survival. Thus adjustment for l is needed and one should rather use

$$p(1_s \,||\, 1_a, 1_b) \sim \sum_{x_l} \hat{p}(1_s \,|\, x_l, 1_a, 1_b)\hat{p}(x_l).$$

Both answers disagree with the correct calculation as given above.

The calculation is a special case of more general situations described by Robins, where randomized treatment allocations and intermediate responses alternate as $t_1, r_1, t_2, r_2, \ldots, t_k, r_k$ and where, for example, the effect of a combined treatment, fixing $t_1, \ldots, t_k$, on the final response r_k is desired. This is then found by G-computation, involving two steps:

1. modifying the joint distribution of all variables (corresponding to a complete DAG in many cases) by the intervention formula (1.13);
2. calculating the marginal distribution of X_{r_k} by a recursive forward computation, possibly using Monte Carlo methods.

We refrain here from describing the more general situation where the suggested treatment regime is allowed to depend on previous treatments and recordings, but emphasize that this does not create essentially new problems.

1.8 Identifiability of causal effects

This section will be concerned with the problem of identifying the effects of interventions from partial observation of a causal system, expressed in the form of a causal directed Markov field. It is largely based on ideas in Pearl (1993) and Pearl (1995a).

1.8.1 The general problem

Consider as usual a finite set of variables V, one of which is labelled t and designated the treatment variable, and another group of variables are considered to be the response, labelled R. We also assume that there is a directed acyclic graph $\mathcal{D}$ such that the joint distribution of all the variables V is causally Markov with respect to $\mathcal{D}$.

The object of interest is the *causal effect* of t on the group of response variables R, represented by the intervention distribution

$$P(X_R = x_R \,|\, X_t \leftarrow x_t^*) = p(x_R \,||\, x_t^*).$$

The remaining variables are partitioned into $C \subseteq V$ and $U \subseteq V$, where C is a set of observed *covariates* whereas the variables in U are to remain unobserved. In principle one could discuss multiple treatments and responses, but this will not be done here.

Thus from an experimental or observational study we obtain information about the joint distribution of the observed variables, t, R and C. Ignoring sampling error, can the causal effect of t on R be determined from this information? Or, phrased in another way, which variables C are needed in order to determine this effect? If the causal effect can be determined from the observed distribution, then how can it be calculated, i.e. is there an analogue of the G-computation that gives the correct answer? If the causal effects cannot be precisely determined, can we at least give inequalities that these numbers must satisfy?

To make the discussion precise, we say that C *identifies* the causal effect of t on R if for any pair P_1, P_2 of distributions that are causally Markov with respect to $\mathcal{D}$ it holds that

$$p_1(x_t, x_C, x_R) \equiv p_2(x_t, x_C, x_R) \implies p_1(x_R \,||\, x_t) \equiv p_2(x_R \,||\, x_t).$$

The most basic question above can now be phrased as determining whether a given set covariates C identifies the causal effect of t on r. Clearly, if $C' \supseteq C$ and C identifies the effect of t on r, so does C', so we are interested in minimal sets of identifying covariates.

Generally C will not identify causal effects unless the conditional distributions are identified by the joint distribution. Thus, *throughout this section we will assume that*

$$p(x_t, x_C) > 0 \quad \textit{for all combinations of } x_t \textit{ and } x_C, \tag{1.16}$$

unless we explicitely state otherwise.

1.8.2 Intervention graphs

When the effect of potential intervention are to be discussed, it is convenient to represent these explicitly in the associated graph of the model considered. As also done in Pearl (1993) and Spirtes et al. (1993), this is done through an *intervention graph* $\mathcal{D}'$, which is formed by augmenting each node representing a variable where intervention is contemplated, with an additional parent.

We denote this additional parent of a vertex α by α'. The corresponding random variable $X_{\alpha'}$ is, when no ambiguity results, just denoted by F_α. The variable F_α has state space $\mathcal{X}_\alpha \cup \{\phi\}$ and the conditional distributions of X_α given its parents in the intervention graph are given by

$$p'(x_\alpha \,|\, x_{\mathrm{pa}(\alpha)}, f_\alpha) = \begin{cases} p(x_\alpha \,|\, x_{\mathrm{pa}(\alpha)}) & \text{if } f_\alpha = \phi \\ \delta_{x_\alpha, x_\alpha^*} & \text{if } f_\alpha = x_\alpha^*, \end{cases} \tag{1.17}$$

where δ_{xy} is Kronecker's symbol

$$\delta_{xy} = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise.} \end{cases}$$

A more general setup would let f_α vary in the set of all (randomized) decision policies, but here we only consider the simpler case.

This approach to the representation of causal effects is related to so-called *influence diagrams* (Howard and Matheson 1984, Shachter 1986, Smith 1989, Oliver and Smith 1990) and taking this connection to its consequence gives yet an alternative basis for causal interpretation of graphical models (Heckerman and Shachter 1995).

Each of the variables $F_\alpha, \alpha \in A$, where A is the set of variables for which intervention is contemplated, can be given an arbitrary distribution with positive probability of all states. We then clearly have

$$p(x) = p'(x \,|\, F_\alpha = \phi, \alpha \in A),$$

but it also holds for any subset $B \subseteq A$ that

$$\begin{aligned} p(x \,||\, x_B^*) &= P(X = x \,|\, X_B \leftarrow x_B^*) \\ &= P'(x \,|\, F_\alpha = x_\alpha^*, \alpha \in B, F_\alpha = \phi, \alpha \in B \setminus A), \quad (1.18) \end{aligned}$$

since it follows from Corollary 1.13 that

$$\begin{aligned} &P'(X = x \,|\, F_\alpha \leftarrow x_\alpha^*, \alpha \in B, F_\alpha \leftarrow \phi, \alpha \in B \setminus A) = \\ &\quad P'(X = x \,|\, F_\alpha = x_\alpha^*, \alpha \in B, F_\alpha = \phi, \alpha \in B \setminus A) \end{aligned}$$

because the variables α' do not have parents. The importance of the relation (1.18) is that it gives a simple connection between intervention conditioning in the original graph and conventional conditioning in the intervention graph.

1.8.3 Three inference rules

The operations needed to find groups of identifying covariates typically involve a sequence of operations that gradually transform expressions involving intervention probabilities to expressions involving ordinary conditional probabilities, the latter being in principle accessible by empirical observation.

We are considering the simple case, where intervention at a node t is contemplated, its effect on a group of variables R is studied, in a context where X_A is observed to be x_A. We let $\perp_{\mathcal{D}'}$ denote d-separation in the intervention graph $\mathcal{D}'$ obtained by augmenting $\mathcal{D}$ with an intervention variable t' as an additional parent of t, and possibly other intervention variables, if also other interventions are contemplated, as described above. We then have the following three inference rules:

Neutral observation of X_t:

$$R \perp_{\mathcal{D}'} t \mid A \implies p(x_r \mid x_A, x_t) = p(x_R \mid x_A) \tag{1.19}$$

Neutral intervention at t:

$$R \perp_{\mathcal{D}'} t' \mid A \implies p(x_r \mid x_A \,||\, x_t) = p(x_R \mid x_A) \tag{1.20}$$

Equivalence of observation and intervention at t:

$$R \perp_{\mathcal{D}'} t' \mid \{A, t\} \implies p(x_R \mid x_A \,||\, x_t) = p(x_R \mid x_A, x_t). \tag{1.21}$$

Each of these can be derived from the directed Markov property of P and P' combined with the fact that intervention probabilites can be obtained by appropriate observation conditioning in the intervention graph.

For example, to derive (1.19) we observe that $R \perp_{\mathcal{D}'} t \mid A$ implies $R \perp_{\mathcal{D}} t \mid A$. This holds because all trails from t to R in $\mathcal{D}$ are also trails in $\mathcal{D}'$ and if one is blocked by A in $\mathcal{D}'$, it is also blocked by A in $\mathcal{D}$. Therefore the global directed Markov property for $\mathcal{D}$ entails that

$$R \perp_{\mathcal{D}'} t \mid A \implies R \perp\!\!\!\perp t \mid A,$$

whereby (1.19) follows.

The relations (1.20) and (1.21) follow directly from the fact that intervention conditioning at t in $\mathcal{D}$ is equivalent to observation conditioning at t' in $\mathcal{D}'$. These rules are also direct consequences of Theorem 7.1 of Spirtes et al. (1993)

Although Pearl (1995a) formulates these inference rules somewhat differently, he conjectures that the three inference rules are complete, in the sense that a set of covariates is identifying for the effect of t on R if and only if all terms involving intervention conditioning in the expression for the intervention distribution can be changed to terms involving observational conditioning by successive application of these three rules. We shall see examples of this in the next subsection, where a number of classical concepts from epidemiology will be illustrated.

1.8.4 The back-door formulae

One of the classic conditions for a set of covariates to be identifying is captured in the theorem below, known as the back-door theorem and formula.

As earlier we contemplate the effect of t on a group of variables R and plan to observe these together with a set of covariates C, whereas the remaining variables in the system are unobserved. Also, as above, $\mathcal{D}'$ denotes the intervention graph obtained from $\mathcal{D}$ by augmenting with an intervention variable t' as an additional parent of t and $\perp_{\mathcal{D}'}$ denotes d-separation in $\mathcal{D}'$.

We then have the following theorem, which can also be derived directly from Theorem 7.1 of Spirtes et al. (1993):

Theorem 1.14 (Back-door) *Assume $C \supseteq C_0$, where C_0 satisfies*

(BD1) *The covariates in C_0 are unaffected by an intervention: $C_0 \perp_{\mathcal{D}'} t'$;*

(BD2) *An intervention only affects the response through the treatment itself, as modified by the observed covariates: $R \perp_{\mathcal{D}'} t' \mid C_0 \cup \{t\}$.*

Then C identifies the effect of the treatment t on R as

$$p(x_R \,||\, x_t^*) = \sum_{x_{C_0}} p(x_R \mid x_{C_0}, x_t^*) p(x_{C_0}). \tag{1.22}$$

Proof: The proof is a simple application of the inference rules. If we partition according to X_{C_0} and then apply first (1.21) for $A = C_0$ and then (1.20) for $R = C_0$ and $A = \emptyset$, we get

$$\begin{aligned} p(x_R \,||\, x_t^*) &= \sum_{x_{C_0}} p(x_R \mid x_{C_0} \,||\, x_t^*) p(x_{C_0} \,||\, x_t^*) \\ &= \sum_{x_{C_0}} p(x_R \mid x_{C_0}, x_t^*) p(x_{C_0} \,||\, x_t^*) \\ &= \sum_{x_{C_0}} p(x_R \mid x_{C_0}, x_t^*) p(x_{C_0}). \end{aligned}$$

□

Condition (BD1) might as well have been formulated by demanding that none of the covariates in C_0 are descendants of t. This is a condition which can then be checked in $\mathcal{D}$ rather than $\mathcal{D}'$.

Note that the positivity assumption (1.16) is important for the joint distribution to identify $p(x_R \mid x_{C_0}, x_t^*)$ for all combinations of its arguments.

In the formulation given, the name ‘back-door theorem’ is not obvious. The lemma below clarifies the reason for the name. A *back-door trail* from t to R in $\mathcal{D}$ is a trail from t to R that does not involve an arrow emanating from t, i.e. leaves t through the ‘back door’. Similarly, we let a *front-door trail* from t to R be a trail that begins with an arrow emanating from t.

Lemma 1.15 *If no covariates in C_0 are descendants of t, $r \perp_{\mathcal{D}'} t' \mid C_0 \cup \{t\}$ if and only if all back-door trails from t to R are blocked by C_0 in $\mathcal{D}$.*

Proof: Assume $r \perp_{\mathcal{D}'} t' \mid C_0 \cup \{t\}$. Each trail from t' to R in $\mathcal{D}'$ corresponds uniquely to a trail from t to R in $\mathcal{D}$. Since the descendants of t are identical in $\mathcal{D}$ and $\mathcal{D}^*$ and none of these are in C_0, t is blocking all trails from t' to R in $\mathcal{D}$ that correspond to front-door trails and it is not blocking any trails corresponding to back-door trails. Consequently, these are be blocked by $C_0 \cup t$ in $\mathcal{D}'$ if and only if they are blocked by C_0 in $\mathcal{D}$. □

The condition in Lemma 1.15 is also phrased in terms of the original graph $\mathcal{D}$ rather than the intervention graph.

James Robins (personal communication) gives the following heuristic argument for the criterion: The treatment effect can be identified if, conditionally on C_0, there is no association beyond causation. Removing arrows out of t eliminates causation. One must thus demand that no conditional association remains after these arrows have been removed.

The formula (1.22) is the classical formula which adjusts for covariates that are not affected by the treatment.

Theorem 1.14 has a slightly more general version, extended by a recursive argument.

Theorem 1.16 (Extended back-door) *Assume $C \supseteq C_0$, where C_0 satisfies*

(EBD1) *The effect of the treatment t on the covariates in C_0 is identified by C;*

(BD2) *An intervention only affects the response through the treatment itself, modified by the observed covariates: $R \perp_{\mathcal{D}'} t' \mid C_0 \cup \{t\}$.*

Then C identifies the effect of the treatment t on R as

$$p(x_R \,||\, x_t^*) = \sum_{x_{C_0}} p(x_R \mid x_{C_0}, x_t^*) p(x_{C_0} \,||\, x_t^*). \qquad (1.23)$$

Proof: This is shown exactly as for Theorem 1.14, just omitting the last step in the calculation. □

Confounding

The first situation to be considered in the light of the back-door theorem is the classical case of a *confounder*, which in the current context is defined to be an unobserved quantity that simultaneously affects the treatment and the response. Thus, in a causal graph, a confounder is a common ancestor to the treatment and response. The literature in epidemiology contains a wealth of more or less precise definitions of the term.

This situation is illustrated in Figure 1.6, displaying the corresponding intervention graph and its associated moral graph. The conditional distribution of X_r after intervention at t cannot be determined from the joint distribution of (X_t, X_r).

Randomization

The next example illustrates how randomization overcomes the identification problem caused by the confounder. Instead of just observing (X_t, X_r), the treatment X_t is now allocated by a known random mechanism, possibly depending on an observed covariate X_c, leading to the diagram described in Figure 1.7. The randomization ensures that there is no arrow pointing from u to t, i.e. the treatment X_t is conditionally independent of X_u given

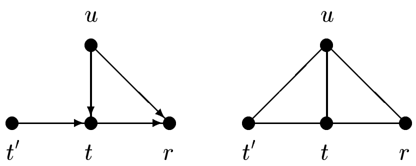


Figure 1.6 *Intervention graph and associated moral graph for experiment with an unobserved confounder. There is a path in the moral graph from t' to r circumventing t so the back-door criterion is violated and the effect of t on r cannot be identified from observations of t and r.*

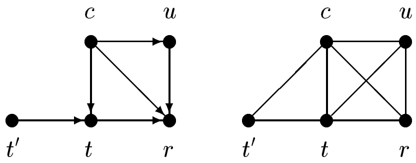


Figure 1.7 *Intervention graph and associated moral graph for experiment with randomized treatment allocation. The moral graph reveals that $r \perp_{\mathcal{D}'} t' \mid \{c, t\}$ so the back-door criterion is satisfied and the treatment effect can be assessed.*

the covariates X_c. To see that (BD1) of Theorem 1.14 is satisfied, we form the ancestral set in $\mathcal{D}'$ generated by $\{c, t'\}$. This is equal to $\{c, t', u\}$ and the associated moral graph has only one edge between c and u. Thus $c \perp_{\mathcal{D}'} t'$. The ancestral set generated by $\{c, t, t', r\}$ is equal to the full set of variables, and the associated moral graph is also displayed in Figure 1.7. Clearly r is separated from t' by $\{c, t\}$ in this graph, so (BD2) is satisfied. Note that if u had been allowed to have an influence on the treatment allocation, the corresponding arrow from u to t in the graph to the left would have induced an edge in the moral graph between t' and u, who were common parents of t, thus violating (BD2) and confounding the relation between t and r.

Sufficient covariate

The next situation to be considered is an observational study where we have no control over the treatment allocation mechanism, but we are able to find a *sufficient* set of covariates, i.e. a set of covariates which is so informative about the response mechanism that the response is conditionally independent of the unobserved variable given the treatment and the covariates. The corresponding intervention graph and associated moral graph is displayed in Figure 1.8.

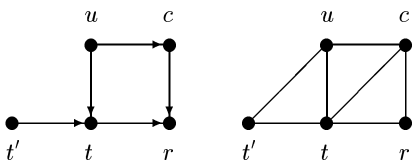


Figure 1.8 *Intervention graph and associated moral graph for an observational study with a sufficient covariate. The moral graph reveals that* $r \perp_{\mathcal{D}'} t' \mid \{c, t\}$ *so the back-door criterion is satisfied and the treatment effect can be assessed.*

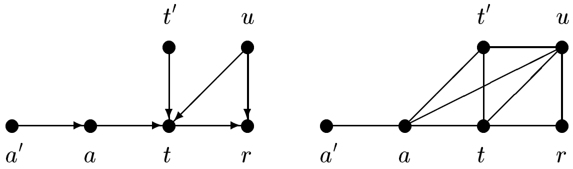


Figure 1.9 *Intervention graph and associated moral graph for a study with partial compliance. The moral graph reveals that* $r \perp_{\mathcal{D}'} a' \mid \{a, t\}$ *so the effect of the treatment assignment is identified. However,* r *is not separated from* t' *by* t *so the effect of the treatment itself cannot be assessed.*

The ancestral set generated by c and t' is equal to $\{c, t', u\}$ and the associated moral graph has only one edge between c and u and thus $c \perp_{\mathcal{D}'} t'$. The ancestral set generated by $\{c, t, t', r\}$ is equal to the full set of variables, and the associated moral graph is displayed in Figure 1.8. Clearly r is separated from t' by $\{c, t\}$ in this graph, so (BD2) is satisfied.

Partial compliance

The next example describes a study in which treatments are assigned completely at random to individuals, but not all individuals are complying with the assignments so that some receive a treatment different from the one assigned. The situation is displayed in Figure 1.9, where a is labelling the assignment and t the actual treatment received. The response, treatment assigned, and treatment received are all observed. From inspection of the moral graph it clearly follows that $r \perp_{\mathcal{D}'} a' \mid \{a, t\}$ so the effect of the treatment assigned is identified via the back-door formula. The corresponding assessment is commonly referred to as “analysis by intention-to-treat”.

However, r is not separated from t' by t so the effect of the treatment itself is not identifiable from these observations. We shall later see how to derive bounds for the effects in this particular case.

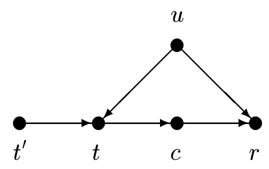


Figure 1.10 *Intervention graph associated with a situation where Theorem 1.17 applies. The covariate c is capturing the way in which t is affecting r, possibly modified by the unobserved variable u.*

1.8.5 The front-door formula

Theorem 1.17 below describes yet another situation where the causal effect of a treatment can be identified. Here the observed covariates are to be considered as the active agent determining the response. A basic example to have in mind could be the effect of smoking on lung cancer, the active agent being the tar content in the lungs. There could be an unobserved, say genetic, feature that influenced both the response and the tendency to smoke. The corresponding diagram is displayed in Figure 1.10 and the conditions in the Theorem 1.17 reflect conditional independence relations following from the directed Markov property of this diagram.

Theorem 1.17 (Front-door formula) *Assume that $C \supseteq C_0$, and there is a subset $D \subseteq V \setminus (C_0 \cup R \cup \{t\})$ such that*

(FD1) *The variables in D are unaffected by an intervention: $D \perp_{\mathcal{D}'} t'$;*

(FD2) *An intervention only affects the covariate through the treatment itself, independently of the variables in D: $C_0 \perp_{\mathcal{D}'} (D \cup \{t'\}) \mid t$;*

(FD3) *An intervention only affects the response through the covariate, as modified by the variables in D: $R \perp_{\mathcal{D}'} t' \mid C_0 \cup D$.*

Then C identifies the effect of the treatment t on R as

$$p(x_R \,||\, x_t^*) = \sum_{x_{C_0}} p(x_{C_0} \mid x_t^*) \sum_{x_t} p(x_R \mid x_{C_0}, x_t) p(x_t). \tag{1.24}$$

Proof: We assume without loss of generality that $C = C_0$ and $D = U$. Then we have

$$\begin{aligned} p(x_R \,||\, x_t^*) &= \sum_{x_C, x_U} p(x_R \mid x_C, x_U \,||\, x_t^*) p(x_C, x_U \,||\, x_t^*) \\ &= \sum_{x_C, x_U} p(x_R \mid x_C, x_U) p(x_C \mid x_U \,||\, x_t^*) p(x_U \,||\, x_t^*), \end{aligned}$$

where we have used (FD3) together with (1.20) to deduce that the intervention in the first term of the product is neutral.

Next, (FD2) combined with the semi-graphoid properties (C3') and (C2') (which are satisfied by the relation $\perp_{\mathcal{D}'}$ by Lemma 1.10) yields

$$C \perp_{\mathcal{D}'} t' \,|\, (U \cup \{t\}) \text{ and } C \perp_{\mathcal{D}'} U \,|\, t.$$

Thus (1.21) yields that intervention in the second term may be substituted with ordinary conditioning, and (1.19) that conditioning with x_U then can be ignored.

Further, (FD1) with (1.20) gives that the third term in the product is independent of x_t^* so that we have

$$p(x_R \,||\, x_t^*) = \sum_{x_U, x_C} p(x_R \,|\, x_U, x_C) p(x_C \,|\, x_t^*) p(x_U).$$

If we now rewrite $p(x_U)$ by partioning according to x_t and note that (FD3) with (1.19) allows further conditioning on x_t in the first term, we get

$$\begin{aligned} p(x_R \,||\, x_t^*) &= \sum_{x_U, x_C, x_t} p(x_R \,|\, x_U, x_C, x_t) p(x_C \,|\, x_t^*) p(x_U \,|\, x_t) p(x_t) \\ &= \sum_{x_C} p(x_C \,|\, x_t^*) \sum_{x_t} p(x_R \,|\, x_C, x_t) p(x_t) \end{aligned}$$

and the proof is complete. □

Both the formula and its name are due to Pearl (1995a), where it is given as part of Theorem 2. Pearl's front-door conditions are formulated rather differently and it is not obvious that they are equivalent to those given here, but we believe them to be. In Pearl's formulation (and our terminology), the conditions are:

(FD1') All directed paths from t to R intersect C_0;

(FD2') All trails from t to C_0 in $\mathcal{D}$ contain an arrow out of t;

(FD3') All back-door trails from C_0 to R are blocked by t in $\mathcal{D}$.

The justification for the name is not so obvious, neither in Pearl's formulation nor in the formulation given here. Although it is true that (FD1) and (FD3) together reflect that C_0 blocks front-door paths from t to R, (FD2) rather reflects that back-door paths from t to C_0 are blocked.

Again we note the importance of the positivity assumption (1.16). Without this, we could always take $c = t$ and satisfy all conditions in Theorem 1.17. However, then the necessary conditional distributions would not be identified by the marginal distributions.

1.8.6 Additional examples

We conclude the section on identifiability of treatment effects by discussing a number of additional examples, illustrating the potential use of the front- and back-door formulae.

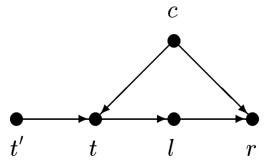


Figure 1.11 *Intervention graph associated with an example where both of the front-door and back-door formulae apply when l and c are observed.*

Example 1.18 Consider the example with intervention graph displayed in Figure 1.11. In the case where c and l are observed together with the treatment t and the response r, the back-door formula applies with c as the covariate and the front-door formula with l as the covariate. The extended back-door theorem applies with both covariates observed. Thus it holds that

$$\begin{aligned} p(x_r \,||\, x_t^*) &= \sum_{x_c} p(x_r \,|\, x_c, x_t^*) p(x_c) \\ &= \sum_{x_l} p(x_l \,|\, x_t^*) \sum_{x_t} p(x_r \,|\, x_l, x_t) p(x_t) \\ &= \sum_{x_c, x_l} p(x_r \,|\, x_l, x_c) p(x_c) p(x_l \,|\, x_t^*). \end{aligned}$$

However, although the treatment effect can be identified in all three observational situations, it is not true that the corresponding maximum likelihood estimates are equally efficient in the case where these are estimated from data. There is clearly loss of information associated with not observing all four variables.

To illustrate this, assume that all variables are discrete and a potential sample of n independent and identically distributed cases with counts $n(x_t, x_r, x_l, x_c)$ are observed, and contrast this with the corresponding incomplete samples only giving $n(x_t, x_r, x_c)$ or $n(x_t, x_r, x_l)$.

In the first situation, the maximum likelihood estimate in the model which is only restricted by satisfying the directed Markov property on the graph is equal to

$$\begin{aligned} \hat{p}(x_t, x_r, x_l, x_c) &= \frac{n(x_c) n(x_t, x_c) n(x_t, x_l) n(x_r, x_l, x_c)}{n\, n(x_c) n(x_t) n(x_l, x_c)} \\ &= \frac{n(x_t, x_c) n(x_t, x_l) n(x_r, x_l, x_c)}{n\, n(x_t) n(x_l, x_c)}, \end{aligned}$$

as each of the conditional probabilities of a variable given its parents is estimated by the corresponding observed relative frequencies (Lauritzen

1996, Theorem 4.36). Using the extended back-door formula, i.e. the last relation above, we therefore get

$$\begin{aligned}\hat{p}(x_r \,||\, x_t^*) &= \sum_{x_c, x_l} \hat{p}(x_r \,|\, x_l, x_c)\hat{p}(x_c)\hat{p}(x_l \,|\, x_t^*) \\ &= \sum_{x_c, x_l} \frac{n(x_r, x_l, x_c)}{n(x_l, x_c)} \frac{n(x_c)}{n} \frac{n(x_l, x_t^*)}{n(x_t^*)}.\end{aligned}$$

The similar expression in the back-door case, i.e. when only $n(x_r, x_c, x_t)$ is observed, becomes

$$\begin{aligned}\hat{p}(x_r \,||\, x_t^*) &= \sum_{x_c} \hat{p}(x_r \,|\, x_c, x_t^*)\hat{p}(x_c) \\ &= \sum_{x_c} \hat{p}(x_r \,|\, x_c, x_t^*) \frac{n(x_c)}{n}.\end{aligned}$$

Note that in this case it is not generally true that we have

$$\hat{p}(x_r \,|\, x_c, x_t^*) = \frac{n(x_r, x_c, x_t^*)}{n(x_c, x_t^*)}$$

because the model induces restrictions on this conditional probability. However, it is obvious that

$$\tilde{p}(x_r \,||\, x_t^*) = \sum_{x_c} \frac{n(x_r, x_c, x_t^*)}{n(x_c, x_t^*)} \frac{n(x_c)}{n}$$

is still a reasonable estimate of the intervention probability. The latter estimate is also the traditional estimate used, and it also applies in the more general case, where the conditional independence $c \perp\!\!\!\perp l \,|\, t$ is violated. Presumably this estimate will be less efficient if indeed the condition $c \perp\!\!\!\perp l \,|\, t$ were known to hold.

In the front-door case, when $n(x_r, x_l, x_t)$ is observed, we similarly have

$$\begin{aligned}\hat{p}(x_r \,||\, x_t^*) &= \sum_{x_l} \hat{p}(x_l \,|\, x_t^*) \sum_{x_t} \hat{p}(x_r \,|\, x_l, x_t)\hat{p}(x_t) \\ &= \sum_{x_l} \frac{n(x_l, x_t^*)}{n(x_t^*)} \sum_{x_t} \hat{p}(x_r \,|\, x_l, x_t) \frac{n(x_t)}{n},\end{aligned}$$

where again the maximum likelihood estimate of the second conditional probability may not be equal to the corresponding relative frequency. However, it is obvious that a reasonable estimate of the treatment effect is equal to

$$\tilde{p}(x_r \,||\, x_t^*) = \sum_{x_l} \frac{n(x_l, x_t^*)}{n(x_t^*)} \sum_{x_t} \frac{n(x_r, x_l, x_t)}{n(x_l, x_t)} \frac{n(x_t)}{n}.$$

It would be interesting to compare the loss of efficiency by not observing c vs. not observing l. □

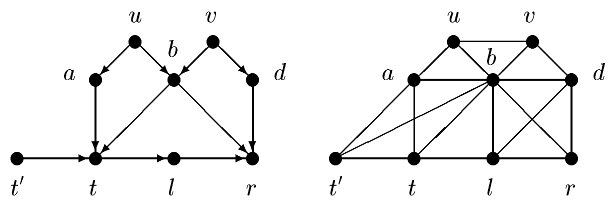


Figure 1.12 *Intervention graph associated with Example 1.19 with its corresponding moral graph.*

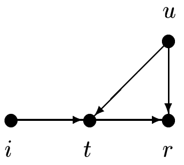


Figure 1.13 *Graphical model expressing that i is an instrumental variable. Note the similarity with the situation of partial compliance described in Figure 1.9, where the assignment variable a is an instrument.*

Example 1.19 The next example is taken from Pearl (1995a) and is somewhat more complex. It is illustrated in Figure 1.12.

As the figure shows, l suffices as a covariate for the front-door formula in Theorem 1.17 to apply.

From the moral graph it is seen by direct inspection that observing b is necessary but not sufficient to satisfy the back-door Theorem 1.14. It needs to be supplemented with any non-empty subset of the variables a, u, v, and d for its union with t to separate t' from r in this graph.

If, for example b, d, and l are observed together with t and r, the extended back-door formula (1.23) yields that the treatment effect is to be estimated from complete data counts as

$$\hat{p}(x_r \,||\, x_t^*) = \sum_{x_b, x_l, x_d} \frac{n(x_r, x_b, x_d, x_l) n(x_l, x_t^*)}{n(x_b, x_d, x_l) n(x_t^*)} \hat{p}(x_b, x_d),$$

where the latter probability ideally should be estimated by taking into account the relevant restrictions induced by the model, rather than using the empirical relative frequencies directly. We leave it to the reader to consider estimation of treatment effects under different observational schemes. □

Figure 1.13 displays the situation in which the variable i is an *instrumental variable* or *instrument* for assessing the effect of t on r. This notion is important in econometrics (Bowden and Turkington 1984, Angrist, Imbens and Rubin 1996). An instrumental variable is one which affects the treatment, but is uncorrelated with unobserved factors. An instrumental variable can be used to derive bounds for treatment effects as we shall show in Section 1.10.1 below.

But here we show an inequality which provides a good example of the restrictions that conditional independence constraints imply for marginal distributions. More precisely, it holds for any discrete treatment variable t that if the independence assumptions associated with the diagram in Figure 1.13 hold, then

$$\sup_{x_t} \int_{x_r} \sup_{x_i} f(x_r, x_t \mid x_i)\, \mu_r(dx_r) \leq 1, \tag{1.25}$$

where f is a generic symbol for the appropriate (conditional) density. This *instrumental inequality* was apparently first derived by Pearl (1995b) and we give the (quite simple) proof below.

The conditional independence restrictions imply that

$$f(x_r, x_t \mid x_i) = \int_{x_u} p(x_t \mid x_i, x_u) f(x_r \mid x_t, x_u)\, P_u(dx_u), \tag{1.26}$$

where P_u denotes the marginal distribution of X_u and the remaining entities are appropriate densities. Since the treatment variable is discrete, we have

$$p(x_t \mid x_i, x_u) \leq 1$$

and this must also hold for its supremum

$$h(x_t, x_u) = \sup_{x_i} p(x_t \mid x_i, x_u) \leq 1.$$

Now we get from (1.26) that

$$\begin{aligned} \sup_{x_i} f(x_r, x_t \mid x_i) &= \int_{x_u} h(x_t, x_u) f(x_r \mid x_t, x_u)\, P_u(dx_u) \\ &\leq \int_{x_u} f(x_r \mid x_t, x_u)\, P_u(dx_u), \end{aligned}$$

whereby

$$\begin{aligned} \int_{x_r} \sup_{x_i} f(x_r, x_t \mid x_i)\, \mu_r(dx_r) &\leq \int_{x_r} \int_{x_u} f(x_r \mid x_t, x_u)\, P_u(dx_u) \mu_r(dx_r) \\ &= \int_{x_u} \int_{x_r} f(x_r \mid x_t, x_u)\, \mu_r(dx_r) P_u(dx_u) \\ &= 1, \end{aligned}$$

and (1.25) follows.

The importance of the inequality (1.25) is that it makes the assumption that i is an instrument *falsifiable* from observations of (X_i, X_t, X_r).

Note that the discreteness of the variable t is used at a very critical point in the proof of (1.25). At the time of writing it is not known whether the assumption of i being an instrument is falsifiable in the general case. In other words, given an arbitrary joint distribution Q of variables (X_i, X_t, X_r), does there exist a random variable X_u and a distribution P of (X_u, X_i, X_t, X_r) which is Markov with respect to the graph in Figure 1.13 and has Q as its marginal to (X_i, X_t, X_r)? In the case where Q is multivariate Gaussian, the answer to the last question is known to be positive, i.e. such a distribution always exists and 'instrumentality' is therefore not falsifiable in the Gaussian case.

1.9 Structural equation models

As mentioned in Section 1.6, the assumption that the intervention formula (1.13) applies is an additional model assumption that does not follow from the basic axioms of probability. There are different ways of justifying this assumption and in any given context, subject matter knowledge must play an essential rôle in this justification process.

A particular modelling formulation, leading to causal Markov models, has documented its relevance in several areas of application. Structural equation models (Bollen 1989) were invented in the context of genetics (Wright 1921, 1923, 1934) , and exploited in economics (Haavelmo 1943, Wold 1954) and social sciences (Goldberger 1972), see for example Pearl (1998) and Spirtes, Richardson, Meek, Scheines and Glymour (1998) for further discussion.

They were used as the main justification and motivation for studying causal Markov models in Kiiveri, Speed and Carlin (1984) and Kiiveri and Speed (1982), as well as in Pearl (1995a) and Pearl (2000).

Most commonly, structural equation models have been assumed linear although there are important exceptions (Goldfeld and Quandt 1972). Here we consider a general structural equation system associated with a directed acyclic graph $\mathcal{D}$. More precisely we consider a system of 'equations'

$$X_v \leftarrow g_v(X_{\mathrm{pa}(v)}, U_v), v \in V, \tag{1.27}$$

where the assignments have to be carried out sequentially, in a well-ordering of the directed acyclic graph $\mathcal{D}$, so that at all times, when X_v is about to be assigned a value, all variables in $\mathrm{pa}(v)$ have already been assigned a value.

The variables $U_v, v \in V$ are assumed to be independent. In the literature, correlation is generally allowed among the 'disturbances' U_v. Also non-recursive systems are often studied. Such systems do not correspond to directed Markov models and they are not studied here. Conditional independence properties for cyclic linear structural equation systems have been studied, for example, by Spirtes (1995), Richardson (1996), Spirtes et al. (1998), and Koster (1996, 1999a, 1999b).

The term 'structural equation system' is really misplaced, and 'structural assignment system' would have been much more appropriate. Much controversy in the literature, in particular concerning calculation of intervention effects, is due to treating the assignment systems as equation systems, 'solving' them and uncritically moving variables between the right-hand side and the left-hand side of (1.27). In particular, this matters when interventions are considered.

It is an important aspect of structural equation models that they also specify the way in which intervention is to be carried out. As is implicit in much literature and, for example, quite explicit in Strotz and Wold (1960), the effect of the intervention $X_a \leftarrow x_a^*$ on a variable with label a is simply that the corresponding line in (1.27) is replaced with the assignment described by the intervention. We refer to this process as *intervention by replacement.* Clearly, the justification that this is a reasonable assumption in any given context is no less difficult than the direct justification of the causal Markov assumption, since the latter follows from (1.27), as stated formally below.

Theorem 1.20 *Let $X = (X_v)_{v \in V}$ be determined by a structural equation system corresponding to a given directed acyclic graph $\mathcal{D}$ and let P denote its distribution. If intervention is carried out by replacement, then P is causally Markov with respect to $\mathcal{D}$.*

Proof: Let the vertices of $\mathcal{D}$ be well-ordered as $v_1, \ldots, v_n$ so that the assignments in (1.27) are carried out in the corresponding order. As the variables U_{v_i} are assumed independent, we clearly have

$$U_{v_i} \perp\!\!\!\perp (X_{v_1}, \ldots, X_{v_{i-1}})$$

and thus, from (C2) and (C3)

$$U_v \perp\!\!\!\perp X_{\mathrm{pr}(v)} \mid X_{\mathrm{pa}(v)}.$$

Using (1.27) with (C2) gives

$$X_v \perp\!\!\!\perp X_{\mathrm{pr}(v)} \mid X_{\mathrm{pa}(v)},$$

i.e. the distribution of X satisfies the ordered directed Markov property (DO). Theorem 1.11 now yields that P is directed Markov on $\mathcal{D}$.

As intervention in a structural equation system is made by replacement, it is clear that all conditional distributions except those involving interventions are preserved. Hence the intervention formula (1.13) applies. □

Note that neither the functions g_v nor the random discturbances U_v are uniquely determined from the distribution P, and not even if P is known to be causally Markov. Thus assuming a specific structural equation model (1.27) is generally stronger — in a way which is typically not empirically testable — than just assuming the causal Markov property, as captured in (1.13).

Some authors seem to prefer to use a structural equation model as justification for the causal Markov property, rather than taking this property as a primitive assumption that must stand usual scientific testing. In view of the above, this may not be reasonable unless specific subject matter knowledge naturally leads to such equations.

1.10 Potential responses and counterfactuals

As mentioned, any causal Markov model for a given DAG $\mathcal{D}$ can be represented by a structural equation system, although this can be done in many different ways.

One type of representation deserves particular attention. Observe first that in each of the equations in (1.27), the values of U_v do not matter beyond what they prescribe as values for g_v, for each fixed value of possible parent configurations $x_{\mathrm{pa}(v)}$. Taking this to its consequence, we can introduce maps ω_v

$$\omega_v : \mathcal{X}_{\mathrm{pa}(v)} \to \mathcal{X}_v.$$

Then each pair (g_v, u_v) in (1.27) determines a map ω_v as

$$\omega_v(x_{\mathrm{pa}(v)}) = g_v(x_{\mathrm{pa}(v)}, u_v)$$

and, conversely, for each set of maps ω_v, we can define g_v as

$$g_v(x_{\mathrm{pa}(v)}, \omega_v) = \omega_v(x_{\mathrm{pa}(v)}).$$

Denoting a random map by Ω_v, we can thus define a structural equation system by

$$X_v \leftarrow g_v(X_{\mathrm{pa}(v)}, \Omega_v), v \in V,$$

and such a system is said to have *canonical form.* The random variables $\Omega_v(x_{\mathrm{pa}(v)})$ describes the *potential response*, i.e. the value of X_v that would have been observed, had the parent configuration been equal to $x_{\mathrm{pa}(v)}$. In this sense, the sets of random variables

$$\{\Omega_v(x_{\mathrm{pa}(v)}) : x_{\mathrm{pa}(v)} \in \mathcal{X}_{\mathrm{pa}(v)}\}$$

are *counterfactual.* The variables Ω_v were called *mapping variables* by Heckerman and Shachter (1995).

This approach to causal inference was for example used by Neyman (1923), Rubin (1974, 1978), and Holland (1986), and it plays a fundamental rôle in the methods developed by Robins (1996, 1997), although it is usually introduced in a slightly different context. Counterfactual objects have at all times been at the basis for causal reasoning (Lewis 1973).

Note that in the formulation given above, the variables Ω_v are no more and no less counterfactual than the ω used when a random variable X is considered to be a deterministic function $X(\omega)$ of a random element ω. This has proved useful in many contexts, although it has also lead to paradoxes, when consequences have been taken too far.

Dawid (2000) argues strongly against the use of counterfactual random variables as for any given individual it is impossible to observe more than one of the variables $\Omega_v(x_{\mathrm{pa}(v)})$; the counterfactual variables are *complementary*. Thus it is dangerous to make assumptions concerning the joint distribution of $\{\Omega_v(x_{\mathrm{pa}(v)}) : x_{\mathrm{pa}(v)} \in \mathcal{X}_{x_{\mathrm{pa}(v)}}\}$, as such distributions are purely metaphysical. And, as it seems that all interesting results concerning causal inference can be derived without counterfactuals, the pitfalls associated with their use can be avoided.

1.10.1 Partial compliance revisited

In this section we show how to use counterfactual variables to get bounds for treatment effects in the case of partial compliance, corresponding to the situation displayed in Figure 1.9. Although as mentioned, the bound can be derived without using conterfactual random variables, they seem to yield a simple method for deriving these bounds in the present example.

With the same notation as earlier, we are interested in the intervention probabilities

$$p(x_r \,||\, x_t^*) = \int_{x_u} p(x_r \,|\, x_t, x_u)\, P_u(dx_u). \tag{1.28}$$

However, only joint observations of a, t, and r are possible. Assuming that we have an infinite sample, we can observe all combinations of

$$p(x_r, x_t \,|\, x_a) = \int_{x_u} p(x_r \,|\, x_t, x_u) p(x_t \,|\, x_a, x_u)\, P_u(dx_u). \tag{1.29}$$

As neither of the back-door or front-door criterions apply, the treatment effect appears not to be identifiable, but it is possible to derive bounds for the intervention probabilities in (1.28) subject to the 'constraints' given in (1.29).

For simplicity we assume that all observed values are binary taking the values 0 or 1. In this case there are a total of six independent constraints, three for each group of treatment assignment.

Bounds for the probabilities involved can be derived in many ways. For example, the bounds (1.25) derived for instrumental variables apply to the observed frequencies here since a is indeed an instrument. Thus this part of the assumptions can and should be checked with observed data. Bounds for treatment effects were also derived by Robins (1989) and Manski (1990). However, it is not always easy to check that the bounds derived are sharp and indeed Balke and Pearl (1994, 1997) derive sharper bounds and show that the bounds cannot be improved. Their argument is based upon the use of counterfactual variables and we shall sketch their argument below.

It may be illuminating to phrase the arguments in terms of the example also considered by Imbens and Rubin (1997) and Balke and Pearl (1997). The example considered is thus the study of the effects on child mortality of vitamin A supplementation in Sumatra, as described by Sommer, Tarwotjo, Djunaedi, West, Leodin, Tilden and Mele (1986) and Sommer and Zeger (1991).

Also here the first part of the argument is that it is not the value or nature of X_u that matters, but only the way in which it affects the two responses t and r. Thus — as was also done by Imbens and Rubin (1997) — we can without loss of generality assume that the unobserved variable is the pair of *potential responses* $\omega = (\omega_t, \omega_r)$, where $\omega_t(x_a)$ denotes the treatment taken by an individual with assigned treatment x_a, and $\omega_r(x_t)$ indicates the response of an individual with treatment x_t.

Each of the potential response variables varies in a space of four elements, so the unobserved variable ω has a total of 16 possible values. The four values of the first variable ω_t may well be called

$$\{\textit{always taker}, \textit{never taker}, \textit{complier}, \textit{defier}\},$$

so that we have *always taker* $(x_a) = 1$, where 1 denotes that vitamin A is taken, *complier* $(x_a) = x_a$ etc. Similarly the four values of ω_t may be called

$$\{\textit{always cured}, \textit{never cured}, \textit{beneficial}, \textit{damaging}\}.$$

In these terms we can rewrite the equations (1.28) and (1.29) as

$$p(x_r \,||\, x_t^*) = \sum_{\omega} p(x_r \,|\, x_t, \omega) p(\omega). \tag{1.30}$$

and

$$p(x_r, x_t \,|\, x_a) = \sum_{\omega} p(x_r \,|\, x_t, \omega) p(x_t \,|\, x_a, \omega) p(\omega). \tag{1.31}$$

The difference between these and those above are that the conditional probabilities in (1.30) and (1.31) are known and equal to one or zero. Thus the problem of finding bounds can be solved by linear programming methods that also identify the best possible bounds. If we let $p_{ij.k} = p(i_r, j_t \,|\, k_a)$ and $q_{ij} = p(i_r \,||\, j_a)$, the bounds were found to be

$$\left.\begin{array}{c} p_{10.1} \\ p_{01.0} \\ p_{10.0} + p_{11.0} - p_{00.1} - p_{11.1} \\ p_{01.0} + p_{10.0} - p_{00.1} - p_{01.1} \end{array}\right\} \leq q_{10} \leq \left\{\begin{array}{c} 1 - p_{00.1} \\ 1 - p_{00.0} \\ p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1} \\ p_{10.0} + p_{11.0} + p_{01.1} + p_{10.1} \end{array}\right.$$

and the remaining bounds are obtained by suitable index substitution.

The bounds turn out to be quite wide in the example mentioned and thus the analysis is inconclusive in this case. Imbens and Rubin (1997), make a full Bayesian analysis of the model, by imposing prior assumptions on the distribution of the potential responses, and thereby obtains the conclusion that the effect of vitamin A is beneficial on average. However, such prior assumptions are untestable and may therefore be questionable. See also Chickering and Pearl (1999) for a further discussion of this example.

As demonstrated in Balke and Pearl (1997), the bounds are sometimes tight and sharp conclusions therefore available. This holds for example for data concerning lipids and coronary heart disease analysed by Efron and Feldman (1991).

1.11 Other issues

1.11.1 Extension to chain graphs

The intervention calculus can be extended to more general graphical models than those given by directed acyclic graphs. Chain graph models are given by graphs that have both directed and undirected links, but no cycles that can be traversed only in one direction without going against the arrows.

The *chain components* $\mathcal{T}$ of such graphs are undirected graphs that are obtained by removing all directed arrows from a chain graph. They naturally unify directed acyclic graphs and undirected graphs in that undirected graphs are chain graphs with only one chain component, and directed acyclic graphs are chain graphs with all chain components being singletons. There is a corresponding set of Markov properties associated with chain graphs (Frydenberg 1990a, Lauritzen 1996). In terms of factorization, the chain graph Markov property manifests itself through an outer factorization

$$f(x) = \prod_{\tau \in \mathcal{T}} f\left(x_\tau \,|\, x_{\mathrm{pa}(\tau)}\right), \tag{1.32}$$

where each factor further factorizes according to the graph $\mathcal{G}^*(\tau)$ as

$$f\left(x_\tau \,|\, x_{\mathrm{pa}(\tau)}\right) = Z^{-1}\left(x_{\mathrm{pa}(\tau)}\right) \prod_{A \in \mathcal{A}(\tau)} \phi_A(x_A), \tag{1.33}$$

where $\mathcal{A}(\tau)$ are the complete sets in $\mathcal{G}^*(\tau)$ and

$$Z\left(x_{\mathrm{pa}(\tau)}\right) = \sum_{x_\tau} \prod_{A \in \mathcal{A}(\tau)} \phi_A(x_A).$$

The graph $\mathcal{G}^*(\tau)$ is obtained from $\mathcal{G}_{\tau \cup \mathrm{pa}(\tau)}$ by dropping directions on edges and adding edges between any pair of members of $\mathrm{pa}(\tau)$.

If the intervention $X_\alpha \leftarrow x^*_\alpha$ is made, the corresponding intervention formula can be argued to be

$$p(x \,||\, x^*_\alpha) = \left.\frac{p(x)}{\sum_{y_{\tau_\alpha} : y_\alpha = x^*_\alpha} p(y_{\tau_\alpha} \,|\, x_{\mathrm{pa}(\tau_\alpha)})}\right|_{x_\alpha = x^*_\alpha} \tag{1.34}$$

where τ_α is the chain component including α. This formula specializes to (1.13) in the fully directed case and (1.14) in the undirected case. This intervention formula corresponds to the analogy with decision networks based on chain graphs as discussed in Cowell et al. (1999). Lauritzen and Richardson (2000) are investigating dynamic regimes that lead to such an intervention calculus and their potential use as an alternative interpretation of simultaneous equation systems.

1.11.2 Causal discovery

Another and more controversial aspect of causal inference from graphical models is associated with identifying causal relationships from data. Ever since the appearance of Glymour, Scheines, Spirtes and Kelly (1987) and the first version of the corresponding program TETRAD, this has been the subject of sometimes quite heated discussions (Freedman 1997, Humphreys and Freedman 1996, Robins and Wasserman 1999, Glymour, Spirtes and Richardson 1999, Humphreys and Freedman 2000).

Basically there have been two different types of approach. The constraint-based approach (Spirtes et al. 1993) is generally conceived to take place in an ideal environment where the joint distribution P of a system X of random variables is known completely without error, whereas the causal graph $\mathcal{D}$ which has generated the distribution is unknown.

Apart from the assumption that such a causal directed acyclic graph $\mathcal{D}$ exists, it is also assumed that P is *faithful* to $\mathcal{D}$, in other words there are no conditional independence relationships between the variables that do not follow from the directed Markov property:

$$A \perp\!\!\!\perp B \mid S \implies A \perp_{\mathcal{D}} B \mid S.$$

As previously mentioned, results of Meek (1995) indicate that most distributions are indeed faithful.

On the assumption above, Spirtes et al. (1993) provide several algorithms that from a relatively modest number of tests identifies the causal graph up to Markov equivalence, i.e. produce a graph $\mathcal{D}'$ with the property that for all disjoint subsets A, B, and S of V

$$A \perp_{\mathcal{D}'} B \mid S \iff A \perp_{\mathcal{D}} B \mid S \iff A \perp\!\!\!\perp B \mid S.$$

They also give variants of these algorithms that do not assume the entire system of variables to be observed. These results are supplemented with conditions for identifiability of causal effects and give methods for identifying causal effects that remain invariant over such an equivalence class.

Richardson and Spirtes (1999) extend the approach to situations involving feedback.

Little has been done to explore the statistical properties of these and similar methods applied to cases where knowledge about the distribution of X is only obtained through finite samples. Although Spirtes et al. (1993) contains a small simulation study, this area deserves to be better explored.

Another line of this research is based on a pure Bayesian approach to learning the structure of a Bayesian network, as initiated by Cooper and Herskovits (1992) and Heckerman, Geiger and Chickering (1995). This approach has been further pursued by Heckerman, Meek and Cooper (1999).

See Cooper (1999) for an overview of the current state of the art within this area.

Acknowledgements

The necessary background for the material presented here was acquired while the author was a Fellow at the Center for Advanced Study in Behavioral Sciences, Stanford, California, with financial support partly provided by National Science Foundation, Grant number #SBR-9022192. The work was completed while the author was a Fellow at the Fields Institute for Research in Mathematical Sciences, Toronto, Canada. I am grateful to these institutions for the excellent working conditions they provide.

The research was also supported by the Danish Research Councils through their PIFT programme.

I am indebted to Thomas Richardson for providing detailed and constructive critical comments to earlier versions of this manuscript.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B.: 1996, Identification of causal effects using instrumental variables (with discussion), *Journal of the American Statistical Association* **91**, 444–472.

Balke, A. and Pearl, J.: 1994, Nonparametric bounds on causal effects from partial compliance data, *in* R. L. de Mantaras and D. Poole (eds), *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, California, pp. 46–54.

Balke, A. and Pearl, J.: 1997, Bounds on treatment effects from studies with imperfect compliance, *Journal of the American Statistical Association* **92**, 1171–1176.

Bollen, K. A.: 1989, *Structural Equations with Latent Variables*, John Wiley and Sons, New York.

Bowden, R. J. and Turkington, D. A.: 1984, *Instrumental Variables*, Cambridge University Press, Cambridge, UK.

Box, G. E. P.: 1966, Use and abuse of regression, *Technometrics* **8**, 625–629.

Chickering, D. M. and Pearl, J.: 1999, A clinician's tool for analyzing noncompliance, *in* C. Glymour and G. F. Cooper (eds), *Computation, Causation, and Discovery*, MIT Press, Cambridge, MA, pp. 407–424.

Cooper, G. F.: 1999, An overview of the representation and discovery of causal relationships using Bayesian networks, *in* C. Glymour and G. F. Cooper (eds), *Computation, Causation, and Discovery*, MIT Press, Cambridge, MA, pp. 3–62.

Cooper, G. F. and Herskovits, E.: 1992, A Bayesian method for the induction of probabilistic networks from data, *Machine Learning* **9**, 309–347.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J.: 1999, *Probabilistic Networks and Expert Systems*, Springer-Verlag, New York.

Cox, D. R.: 1984, Design of experiments and regression, *Journal of the Royal Statistical Society, Series A* **147**, 306–315.

Dawid, A. P.: 2000, Causal inference without counterfactuals, *Journal of the American Statistical Association* **95**, to appear.

Efron, B. and Feldman, D.: 1991, Compliance as an explanatory variable in clinical trials, *Journal of the American Statistical Association* **86**, 9–26.

Freedman, D.: 1997, From association to causation via regression, *in* V. McKim and S. Turner (eds), *Causality in Crisis?*, University of Notre Dame Press, pp. 113–182.

Frydenberg, M.: 1990a, The chain graph Markov property, *Scandinavian Journal of Statistics* **17**, 333–353.

Frydenberg, M.: 1990b, Marginalization and collapsibility in graphical interaction models, *Annals of Statistics* **18**, 790–805.

Geiger, D. and Pearl, J.: 1990, On the logic of causal models, *in* R. D. Shachter, T. S. Levitt, L. N. Kanal and J. F. Lemmer (eds), *Uncertainty in Artificial Intelligence IV*, North-Holland, Amsterdam, pp. 136–147.

Glymour, C. and Cooper, G. F.: 1999, *Computation, Causation, and Discovery*, MIT Press, Cambridge, MA.

Glymour, C., Scheines, R., Spirtes, P. and Kelly, K.: 1987, *Discovering Causal Structure*, Academic Press, New York.

Glymour, C., Spirtes, P. and Richardson, T.: 1999, On the possibility of inferring causation from association without background knowledge, *in* C. Glymour and G. F. Cooper (eds), *Computation, Causation, and Discovery*, MIT Press, Cambridge, MA, pp. 323–331.

Goldberger, A. S.: 1972, Structural equation models in the social sciences, *Econometrica* **40**, 979–2001.

Goldfeld, S. M. and Quandt, R. E.: 1972, *Non-linear Methods in Econometrics*, North-Holland, Amsterdam, Netherlands.

Haavelmo, T.: 1943, The statistical implications of a system of simultaneous equations, *Econometrica* **11**, 1–12.

Hammersley, J. M. and Clifford, P. E.: 1971, Markov fields on finite graphs and lattices, Unpublished manuscript.

Heckerman, D. and Shachter, R.: 1995, Decision-theoretic foundations for causal reasoning, *Journal of Artificial Intelligence Research* **3**, 405–430.

Heckerman, D., Geiger, D. and Chickering, D. M.: 1995, Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning* **20**, 197–243.

Heckerman, D., Meek, C. and Cooper, G.: 1999, A Bayesian approach to causal discovery, *in* C. Glymour and G. F. Cooper (eds), *Computation, Causation, and Discovery*, MIT Press, Cambridge, MA, pp. 141–165.

Holland, P.: 1986, Statistics and causal inference, *Journal of the American Statistical Association* **81**, 945–960.

Howard, R. A. and Matheson, J. E.: 1984, Influence diagrams, *in* R. A. Howard and J. E. Matheson (eds), *Readings in the Principles and Applications of Decision Analysis*, Strategic Decisions Group, Menlo Park, CA.

Humphreys, P. and Freedman, D.: 1996, The grand leap, *British Journal for the Philosophy of Science* **47**, 113–123.

Humphreys, P. and Freedman, D.: 2000, Are there algorithms that discover causal structure?, *Synthese*. In press.

Imbens, G. W. and Rubin, D. B.: 1997, Bayesian inference for causal effects in randomized experiments with noncompliance, *Annals of Statistics* **25**, 305–327.

Jensen, F. V., Lauritzen, S. L. and Olesen, K. G.: 1990, Bayesian updating in causal probabilistic networks by local computation, *Computational Statistics Quarterly* **4**, 269–282.

Kiiveri, H. and Speed, T. P.: 1982, Structural analysis of multivariate data: A review, *in* S. Leinhardt (ed.), *Sociological Methodology*, Jossey-Bass, San Francisco.

Kiiveri, H., Speed, T. P. and Carlin, J. B.: 1984, Recursive causal models, *Journal of the Australian Mathematical Society, Series A* **36**, 30–52.

Koster, J. T. A.: 1996, Markov properties of non-recursive causal models, *Annals of Statistics* **24**, 2148–2177.

Koster, J. T. A.: 1999a, Linear structural equations and graphical models, Lecture Notes. The Fields Institute, Toronto, Canada.

Koster, J. T. A.: 1999b, On the validity of the Markov interpretation of path diagrams of Gaussian structural equation systems with correlated errors, *Scandinavian Journal of Statistics* **26**, 413–431.

Lauritzen, S. L.: 1996, *Graphical Models*, Clarendon Press, Oxford, United Kingdom.

Lauritzen, S. L. and Richardson, T. S.: 2000, Chain graph models for intervention, Manuscript in preparation.

Lauritzen, S. L. and Spiegelhalter, D. J.: 1988, Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *Journal of the Royal Statistical Society, Series B* **50**, 157–224.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer, H.-G.: 1990, Independence properties of directed Markov fields, *Networks* **20**, 491–505.

Lewis, D.: 1973, *Counterfactuals*, Harvard University Press, Cambridge, MA.

Manski, C. F.: 1990, Nonparametric bounds on treatment effects, *American Economic Review, Papers and Proceedings* **80**, 319–323.

Meek, C.: 1995, Strong completeness and faithfulness in Bayesian networks, *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, Morgan Kaufmann Publishers, San Francisco, CA, pp. 411–418.

Neyman, J.: 1923, On the application of probability theory to agricultural experiments. Essay on principles, In Polish. English translation of Section 9 by D. Dabrowska and T. P. Speed in *Statistical Science* **5** (1990), 465–480.

Oliver, R. M. and Smith, J. Q.: 1990, *Influence Diagrams, Belief Nets and Decision Analysis*, John Wiley and Sons, Chichester, United Kingdom.

Pearl, J.: 1986a, A constraint–propagation approach to probabilistic reasoning, *in* L. N. Kanal and J. F. Lemmer (eds), *Uncertainty in Artificial Intelligence*, North-Holland, Amsterdam, The Netherlands, pp. 357–370.

Pearl, J.: 1986b, Fusion, propagation and structuring in belief networks, *Artificial Intelligence* **29**, 241–288.

Pearl, J.: 1988, *Probabilistic Inference in Intelligent Systems*, Morgan Kaufmann Publishers, San Mateo, CA.

Pearl, J.: 1993, Graphical models, causality and intervention, *Statistical Science* **8**, 266–269. Comment to Spiegelhalter *et al.* (1993).

Pearl, J.: 1995a, Causal diagrams for empirical research, *Biometrika* **82**, 669–710.

Pearl, J.: 1995b, Causal inference from indirect experiments, *Artificial Intelligence in Medicine* **7**, 561–582.

Pearl, J.: 1998, Graphs, causality, and structural equation models, *Sociological Methods and Research* **27**, 226–284.

Pearl, J.: 2000, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, UK.

Pearl, J. and Paz, A.: 1987, Graphoids: A graph based logic for reasoning about relevancy relations, *in* B. D. Boulay, D. Hogg and L. Steel (eds), *Advances in Artificial Intelligence—II*, North-Holland, Amsterdam, pp. 357–363.

Richardson, T. and Spirtes, P.: 1999, Automated discovery of linear feedback models, *in* C. Glymour and G. F. Cooper (eds), *Computation, Causation, and Discovery*, MIT Press, Cambridge, MA, pp. 253–304.

Richardson, T. S.: 1996, *Models of Feedback: Interpretation and Discovery*, PhD thesis, Carnegie-Mellon University.

Robins, J. M.: 1986, A new approach to causal inference in mortality studies with sustained exposure periods — application to control of the healthy worker survivor effect, *Mathematical Modelling* **7**, 1393–1512.

Robins, J. M.: 1989, The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, *in* L. Sechrest, H. Freeman, and A. Mulley (eds), *Health Service Research Methodology: A Focus on AIDS*, U.S. Public Health Service, Washington DC, pp. 113–159.

Robins, J. M.: 1997, Causal inference from complex longitudinal data, *in* M. Berkane (ed.), *Latent Variable Modelling and Applications to Causality*, Vol. 120 of *Lecture Notes in Statistics*, Springer-Verlag, New York, pp. 69–117.

Robins, J. M. and Wasserman, L.: 1999, On the impossibility of inferring causation from association without background knowledge, *in* C. Glymour and G. F. Cooper (eds), *Computation, Causation, and Discovery*, MIT Press, Cambridge, MA, pp. 305–321.

Rubin, D. B.: 1974, Estimating causal effects of treatments in randomized and non-randomized studies, *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B.: 1978, Bayesian inference for causal effects. The role of randomization, *Annals of Statistics* **6**, 34–58.

Shachter, R. D.: 1986, Evaluating influence diagrams, *Operations Research* **34**, 871–882.

Shafer, G.: 1985, Conditional probability, *International Statistical Review* **53**, 261–277.

Shafer, G.: 1996, *The Art of Causal Conjecture*, MIT Press, Cambridge, Massachusetts.

Shenoy, P. P. and Shafer, G.: 1990, Axioms for probability and belief–function propagation, *in* R. D. Shachter, T. S. Levitt, L. N. Kanal and J. F. Lemmer (eds), *Uncertainty in Artificial Intelligence 4*, North-Holland, Amsterdam, The Netherlands, pp. 169–198.

Smith, J. Q.: 1989, Influence diagrams for Bayesian decision analysis, *European Journal of Operational Research* **40**, 363–376.

Sommer, A. and Zeger, S. L.: 1991, On estimating efficacy from clinical trials, *Statistics in Medicine* **10**, 45–52.

Sommer, A., Tarwotjo, I., Djunaedi, E., West, K. P., Leodin, A. A., Tilden, R. and Mele, L.: 1986, Impact of vitamin A supplementation on childhood mortality: A randomized controlled community trial, *The Lancet* pp. 1169–1173.

Speed, T. P.: 1979, A note on nearest-neighbour Gibbs and Markov probabilities, *Sankhyā, Series A* **41**, 184–197.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G.: 1993, Bayesian analysis in expert systems (with discussion), *Statistical Science* **8**, 219–283.

Spirtes, P.: 1995, Directed cyclic graphical representations of feedback models, *in* P. Besnard and S. Hanks (eds), *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, Morgan Kaufmann Publishers, San Francisco, CA, pp. 491–498.

Spirtes, P., Glymour, C. and Scheines, R.: 1993, *Causality, Prediction and Search*, Springer-Verlag, New York.

Spirtes, P., Richardson, T., Meek, C., Scheines, R. and Glymour, C.: 1998, Using path diagrams as a structural modelling tool, *Sociological Methods and Research* **27**, 182–225.

Strotz, R. H. and Wold, H. O. A.: 1960, Recursive versus nonrecursive systems: An attempt at synthesis, *Econometrica* **28**, 417–427.

Verma, T. and Pearl, J.: 1990, Causal networks: Semantics and expressiveness, *in* R. D. Shachter, T. S. Levitt, L. N. Kanal and J. F. Lemmer (eds), *Uncertainty in Artificial Intelligence 4*, North-Holland, Amsterdam, The Netherlands, pp. 69–76.

Wold, H. O. A.: 1954, Causality and econometrics, *Econometrica* **22**, 162–177.

Wright, S.: 1921, Correlation and causation, *Journal of Agricultural Research* **20**, 557–585.

Wright, S.: 1923, The theory of path coefficients: a reply to Niles' criticism, *Genetics* **8**, 239–255.

Wright, S.: 1934, The method of path coefficients, *Annals of Mathematical Statistics* **5**, 161–215.