CHAPTER ONE

An Overview of the Representation and Discovery of Causal Relationships Using Bayesian Networks

Gregory F. Cooper

To assist readers who are new to the area of causal discovery from observational data, this chapter describes some important concepts. The emphasis is on informal insight rather than formal rigor. The chapter introduces Bayesian networks as a representation of causal relationships. Other representations of causality are mentioned, but are not discussed in detail. Some basic properties of causal and noncausal Bayesian networks are described. Several fundamental methods and assumptions for learning causal Bayesian networks from observational data are introduced, and strengths and weaknesses of the methods are also discussed.

1. Bayesian Networks

A *Bayesian network* consists of a structural model and a set of probabilities (Castillo, Gutierrez, and Hadi 1997; Jensen 1996; Neapolitan 1990; Pearl 1988; Spirtes, Glymour, and Scheines 1993). The structural model is a directed acyclic graph[1] in which nodes represent variables and arcs represent probabilistic dependence. For convenience, I will use the terms *node* and *variable* interchangeably in this chapter. Each node can represent a continuous or discrete variable. For each node there is a probability distribution on that node given the state of its parents. A Bayesian network specifies graphically how the node probabilities factor to specify a joint probability distribution over all the nodes (variables).

Let S be the graphical structure of a Bayesian network G and let P be the

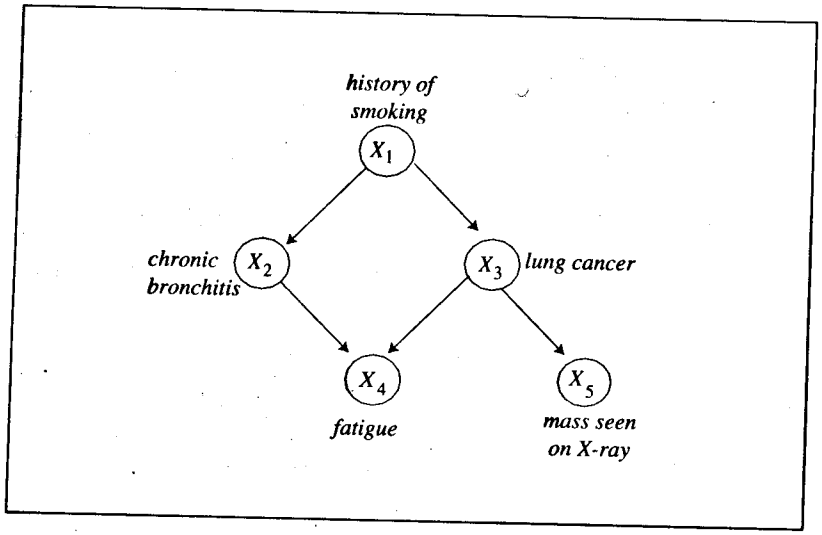

Figure 1. A hypothetical Bayesian network structure.

joint probability distribution represented by G. By definition, S is a directed, acyclic graph. A node in S denotes a variable that models a feature of a process, event, state, object, agent, etc., all of which I will denote generically as an entity. For example, the presence or absence of fatigue is a feature of a patient, and the patient is an entity. S may contain both measured (i.e., recorded) and hidden (i.e., unrecorded) variables. Hidden variables are variables for which we have no data.

Figure 1 illustrates the structure of a hypothetical medical Bayesian network, which contains five nodes. Table 1 shows the probabilities that are associated in this example with the structure in figure 1.

1.1. The Markov Condition

The independence relationships represented by the structure of a Bayesian network are given by the *Markov condition*:

> Any node is conditionally independent of its nondescendants, given its parents.

A descendant of a node X is a node Y that can be reached by a directed path from X to Y. The Markov condition can be used to define equivalence classes of network structures. Two Bayesian network structures are *Markov equivalent* if and only if they contain the same set of variables and they represent the same conditional independence relationships on those variables, as

$P(X_1 = \text{no}) = 0.80$	$P(X_1 = \text{yes}) = 0.20$
$P(X_2 = \text{absent} \mid X_1 = \text{no}) = 0.95$ $P(X_2 = \text{absent} \mid X_1 = \text{yes}) = 0.75$	$P(X_2 = \text{present} \mid X_1 = \text{no}) = 0.05$ $P(X_2 = \text{present} \mid X_1 = \text{yes}) = 0.25$
$P(X_3 = \text{absent} \mid X_1 = \text{no}) = 0.99995$ $P(X_3 = \text{absent} \mid X_1 = \text{yes}) = 0.997$	$P(X_3 = \text{present} \mid X_1 = \text{no}) = 0.00005$ $P(X_3 = \text{present} \mid X_1 = \text{yes}) = 0.003$
$P(X_4 = \text{absent} \mid X_2 = \text{absent}, X_3 = \text{absent}) = 0.95$ $P(X_4 = \text{absent} \mid X_2 = \text{absent}, X_3 = \text{present}) = 0.50$ $P(X_4 = \text{absent} \mid X_2 = \text{present}, X_3 = \text{absent}) = 0.90$ $P(X_4 = \text{absent} \mid X_2 = \text{present}, X_3 = \text{present}) = 0.25$	$P(X_4 = \text{present} \mid X_2 = \text{absent}, X_3 = \text{absent}) = 0.05$ $P(X_4 = \text{present} \mid X_2 = \text{absent}, X_3 = \text{present}) = 0.50$ $P(X_4 = \text{present} \mid X_2 = \text{present}, X_3 = \text{absent}) = 0.10$ $P(X_4 = \text{present} \mid X_2 = \text{present}, X_3 = \text{present}) = 0.75$
$P(X_5 = \text{absent} \mid X_3 = \text{absent}) = 0.98$ $P(X_5 = \text{absent} \mid X_3 = \text{present}) = 0.40$	$P(X_5 = \text{present} \mid X_3 = \text{absent}) = 0.02$ $P(X_5 = \text{present} \mid X_3 = \text{present}) = 0.60$

Table 1. The probabilities associated with figure 1.

These probabilities are for illustration only; they are not intended to accurately reflect frequencies of events in any actual patient population.

given by the Markov condition. For example, consider a two-node Bayesian network. The network structure $X \rightarrow Y$ is Markov equivalent to $X \leftarrow Y$, because both networks represent the same conditional independence relationships between X and Y (namely, none). Neither network is Markov equivalent to a structure with no arc between X and Y, which we will represent as X no_arc Y.

The Markov condition also permits the factorization of a joint probability distribution on model variables $X_1, X_2, \ldots, X_n$ into the following product (Pearl 1988):

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i)) \qquad (1)$$

where $parents(X_i)$ denotes the set of nodes with arcs into X_i. If X_i has no parents, then the set $parents(X_i)$ is empty, and therefore $P(X_i \mid parents(X_i))$ is just $P(X_i)$.

Consider the example given by figure 1. Equation 1 permits the derivation of a joint probability on the five model variables as follows:

$$P(X_1, X_2, X_3, X_4, X_5)$$
$$= P(X_5 \mid X_3)P(X_4 \mid X_2, X_3)P(X_3 \mid X_1)P(X_2 \mid X_1)P(X_1)$$

Thus, for example, $P(X_1 = \text{yes}, X_2 = \text{present}, X_3 = \text{present}, X_4 = \text{present}, X_5 = \text{present}) = 0.20 \times 0.25 \times 0.003 \times 0.75 \times 0.60 = 0.0000675$. The factorization entailed by a Bayesian network often allows a compact representa-

tion of the complete joint probability distribution. For instance, in the previous example, the exhaustively enumerated joint distribution requires 32 probabilities. In contrast, table 1 contains only 11 independent probabilities (the other 11 in table 1 can be derived from the axioms of probability theory). Often, unless a Bayesian network structure contains a high density of arcs, the amount of space savings is substantial.

Consider a node X in a Bayesian network. The *Markov blanket* of X is defined as the set of nodes consisting of the parents of X, the children of X, and the parents of the children of X. It follows from the Markov condition that if we condition on values of each node in the Markov blanket of X, then X is probabilistically independent of all other nodes in the network other than X and its Markov blanket.

1.2. The *d*-Separation Criterion

A graphical criterion called *d-separation* captures exactly all the conditional independence relationships that are implied by the Markov condition (Geiger, Verma, and Pearl 1996; Meek 1995; and Pearl 1988), which was defined in section 1.1. The following is a definition of *d*-separation (Pearl 1995):

> Let A, B, and C be disjoint subsets of the nodes in S. Let p be any acyclic path between a node in A and a node in B, where an acyclic path is any succession of arcs, regardless of their directions, such that no node along those arcs appears more than once. We say a node w has converging arrows along a path if two arcs on the path point to w. Subset C is said to block p if there is a node w on p satisfying one of the following two conditions: (1) w has converging arrows (along p) and neither w nor any of its descendants are in C, or (2) w does not have converging arrows (along p) and w is in C. Subset C is said to *d*-separate A from B in S if and only if C blocks every acyclic path from a node in A to a node in B.

If A and B are not *d*-separated given C, then we say they are *d-connected* given C. For example, in the Bayesian network structure in figure 1, X_2 and X_3 are *d*-separated given X_1, which implies that X_2 and X_3 are conditionally independent given X_1. If we do not condition on X_1, then X_2 and X_3 are not *d*-separated, because the path from X_2 to X_3 through X_1 is not blocked. As another example, X_4 is *d*-separated from X_1 by X_2 and X_3 . As a final example, suppose we remove X_1 from figure 1 to create a new example network, then X_2 and X_3 would be *d*-separated, without any conditioning. Note, however, that once we condition on X_4, then X_2 and X_3 are not *d*-separated; in section 4.5.3 I provide a causal justification for such *d*-connectivity.

1.3. Probabilistic Inference Using a Bayesian Network

In this section, I discuss how a given Bayesian network can be used to derive the posterior probability distribution of one or more variables in the network given that we condition on the values *observed* for other variables in the network. Such inferences presume, of course, that the network has already been constructed, either manually or with the aid of automated learning methods, such as those I describe later in this chapter. While this section focuses on inferences given only observations, in section 2 I discuss how to derive a posterior probability distribution of a variable when we observe some of the variables and *manipulate* other variables.

Since a Bayesian network encodes a joint probability distribution, as given by equation 1, it contains all the information needed to compute any marginal or conditional probability on the nodes in the network. Using the sample network given by figure 1 and table 1, we can derive the following five probabilities (and many more):

$P(X_1 = yes \mid X_4 = present)$
$P(X_1 = yes \mid X_4 = absent \text{ and } X_5 = present)$
$P(X_4 = present \text{ and } X_5 = present \mid X_1 = no)$
$P(X_1 = yes \text{ and } X_4 = absent \mid X_2 = present \text{ and } X_5 = present)$
$P(X_2 = present \text{ and } X_3 = present)$

Let S and T be sets of variables with assigned values. For example, S might be $\{X_1 = yes\}$ and T might be $\{X_4 = \text{present}\}$. Suppose we wish to know $P(S \mid T)$. Conceptually, we can view inference as a simple procedure in which the marginals $P(S \cup T)$ and $P(T)$ are computed, and $P(S \cup T)/P(T)$ is returned. Consider deriving the marginal probability $P(T)$. Let U be the variables in the network that do not appear in set T. Using equation 1, we can sum over all U to obtain $P(T)$ as follows:

$$P(T) = \sum_{U} \prod_{i=1}^{n} P(X_i \mid parents(X_i)) \qquad (2)$$

where the sum is taken over all unique combinations of value assignments to the variables in U, and in the product if X_i appears in T, then X_i is assigned the value given by T. For the example in which T is $\{X_4 = \text{present}\}$, the application of equation 2 yields:

$P(T) =$
$P(X_1 = no)\, P(X_2 = absent \mid X_1 = no)\, P(X_3 = absent \mid X_1 = no)$
$P(X_4 = present \mid X_2 = absent, X_3 = absent)\, P(X_5 = absent \mid X_3 = absent)$
$+ P(X_1 = yes)\, P(X_2 = absent \mid X_1 = yes)\, P(X_3 = absent \mid X_1 = yes)$
$P(X_4 = present \mid X_2 = absent, X_3 = absent)\, P(X_5 = absent \mid X_3 = absent)$

$$+ P(X_1 = no)\, P(X_2 = present \mid X_1 = no)\, P(X_3 = absent \mid X_1 = no)$$
$$P(X_4 = present \mid X_2 = present, X_3 = absent)\, P(X_5 = absent \mid X_3 = absent)$$
$$\vdots$$
$$+ P(X_1 = yes)\, P(X_2 = present \mid X_1 = yes)\, P(X_3 = present \mid X_1 = yes)$$
$$P(X_4 = present \mid X_2 = present, X_3 = present)$$
$$P(X_5 = present \mid X_3 = present)$$

which by equation 1 is equal to

$$P(X_1 = \text{no}, X_2 = absent, X_3 = absent, X_4 = present, X_5 = absent)$$
$$+ P(X_1 = \text{yes}, X_2 = absent, X_3 = absent, X_4 = present, X_5 = absent)$$
$$+ P(X_1 = \text{no}, X_2 = present, X_3 = absent, X_4 = present, X_5 = absent)$$
$$\vdots$$
$$+ P(X_1 = yes, X_2 = present, X_3 = present, X_4 = present, X_5 = present)$$

That is, we sum over every possible joint instantiation of the variables, while holding the variables in T constant to their assigned values. A serious practical problem with using equation 2 is that its time complexity is exponential in the number of variables in U. Thus, often this simple, brute-force inference algorithm is not computationally tractable. Researchers have developed general inference algorithms that can take advantage of independence relationships represented in a Bayesian network to often perform inference much more efficiently than equation 2 (Jensen 1996). Indeed, for some networks, inference can be performed in time that is polynomial in the number of nodes in the network. For example, if a network has only one path between any two nodes, then algorithms have been developed that perform inference in time that is linear in the size of the Bayesian network (Kim and Pearl 1983, Pearl 1988). Nonetheless, it has been shown that inference is NP-hard (Cooper 1990). Thus, we would not expect to find an inference algorithm that is efficient (i.e., polynomial-time in the size of the network) in the worst cases for all Bayesian networks. High computational complexity results from having multiple pathways between nodes in a network. For a network with multiple pathways, typically the number of pathways between nodes increases with the number of arcs, making exact inference more computationally expensive. When exact inference is prohibitively time consuming, stochastic approximation algorithms can be applied (Henrion 1990). These algorithms may yield useful estimates of exact inference results, although in the worst cases stochastic approximation algorithms are unlikely to yield usefully precise estimates (Dagum and Luby 1993).

2. An Operational Test of Causality

The usefulness of causal knowledge stems from its ability to predict how manipulation of the world will (or did) change the world. The immediate goals for acquiring causal knowledge include causal *explanation* of past manipulations and outcomes (e.g., legal liability often is based on the probable causes of an untoward effect), *insight* into the existence of causal mechanisms acting currently (e.g., the side effects caused by a newly introduced drug), and *prediction* of outcomes that will follow from manipulations (e.g., the cure rate of a disease when a particular surgery is performed). In this book, we emphasize insight and prediction.

This book does not attempt to develop a comprehensive, formal definition of causality. Intuitively, however, causal knowledge is knowledge that predicts how actions are likely to change the world. Operationally, for example, we might test for the existence of a causal relationship by using randomized controlled experiments (RCEs), where saying that X *causes* Y means that a hypothetically ideal RCE would conclude that there is some manipulation of X (possibly in concert with the manipulation of other variables; see section 4.3) that leads to a change in the probability distribution of values that Y will take on. Since no claim is being made that such a test can detect all causal relationships, the test is not being proposed here as a definition of causality.

The notion of a *manipulation* is closely related to the concept of an *act* in decision theory (Savage 1954). In most formulations, the application of normative decision theory requires a specification of the probabilities of possible effects of alternative causes. Heckerman and Shachter (1995) develop a formal connection between causality and decision theory.

In what follows, I informally describe a test for causality in terms of manipulations performed within RCEs. I first outline here a prototypical RCE; although variations certainly exist, they are not discussed. An RCE is performed with an explicitly defined population of units (e.g., patients with *chest pain*) in some explicitly defined context or set of contexts (e.g., currently receiving no chest-pain medication and residing in a given geographical area). Thus, causal relationships that are discovered are relative to a population of units and a context. In an RCE, for a given experimental unit, the value to set the cause in question, which I denote as X, is randomly selected using a uniform distribution over the possible values of X. The state of X is then manipulated to have the selected value. The RCE defines explicitly the details of how these manipulations are made (e.g., the quantity of chest-pain medication to take and the time course of taking the medication). For each unit, after the new value of X is set (e.g., either *receive chest-pain medication* or *receive no chest-pain medication*), the value of Y is measured at some designated time later (e.g., either *has chest pains* or *does not have chest*

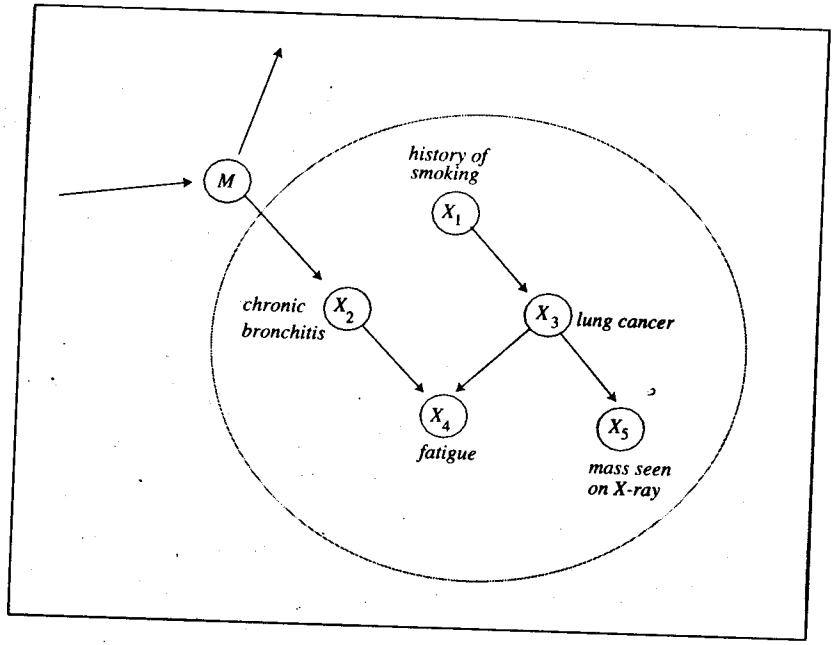

Figure 2. Chronic bronchitis is the variable being manipulated.

pains). The greater the experimental data support a statistical dependency between X and Y, the more the data support that X causally influences Y.

The population units under consideration may have properties, as represented by variable values, that are intrinsic to the meaning of those units. If such values of such a unit were changed as a result of manipulations, then the unit would no longer be in the population under study. For example, if we are interested in causal processes within an intact *e. coli* cell (the population units of interest), but a manipulation completely destroys the cell membrane of *e. coli* cells, then we could not use such a manipulation to study the causal processes of interest.

The notion of a manipulation deserves further description. A manipulation can be represented with a variable that is external to the system being modeled, such that the manipulation variable has an arc to the variable within the system that is being manipulated. We assume that such an external variable is not caused by or confounded with any variable in the modeled system. The marginal probability distribution over the values of the manipulation variable should be positive; that is, it should contain no probabilities of zero or one. With an RCE we can test experimentally that variable X (or variable set $\boldsymbol{X}$) has a causal influence (on one or more other variables) if it is possible for a manipulation variable to influence (i.e., have an arc into) just X (or $\boldsymbol{X}$), without influencing the state of the remaining modeled system variables. The manipulation variable has the same value set as the manipulated variable. The relationship between the two variables is deterministic, so that the manipulated variable always takes on the same value as the manipulation variable; this deterministic linkage ensures that the manipulated variable will not be influenced by any other variable within the modeled system. In essence, manipulation is a special type of causal relationship that we assume exists in order to define and discover more subtle causal relationships within some system of interest.

Figure 2 shows *chronic bronchitis* as the variable being manipulated. The variable named M denotes the manipulation variable, which is outside the encircled system of five modeled variables $X_1, X_2, \ldots, X_5$. Note that the manipulation of X_2 by M means that all other arcs into X_2 (i.e., the arc from X_1 in the example) are causally inconsequential and therefore are removed. The arcs into and out of the manipulation variable M indicate that variables outside the modeled system may causally influence and be influenced by M, as long as they are not causally related to the variables in the modeled system.

The manipulation theorem, which is stated and proved in Spirtes, Glymour, and Scheines (1993) (see also chapters 2 and 3), provides a simple graphical procedure for inferring the posterior probability distribution of variables under manipulation M given observations O. The procedure is as follows: We remove all the arcs into each manipulated variable and set the variable to the value given by the manipulation. We then perform regular Bayesian network inference, as outlined in section 1.3, conditioned on the observations given by O and the instantiated values of the manipulated variables. For example, for the causal network shown in figure 2, suppose we wish to infer the expected distribution of fatigue, given that an individual has a mass seen on X-ray and we cure him or her of any existing bronchitis. In this case, we set *chronic bronchitis* to the value absent and remove its arc from *history of smoking*. We set *mass seen on X-ray* to the value *yes*. We then apply regular Bayesian probabilistic inference methods (see section 1.3) to compute the posterior probability of *fatigue*.

While we can use RCEs to provide (at least conceptually) a test of causality, in practice even a limited RCE might not be safe, ethical, logistically feasible, financially worthwhile, or even theoretically possible, all of which are reasons for using observational data to attempt to infer causal relationships. Because (1) RCEs have limitations and (2) causal discovery from RCEs is well addressed in the literature (see, for example, Bulpitt 1996 and Friedman, Furberg, and DeMets 1996), this book focuses on learning causal relationships from observational data. The ability to use observational data for causal discovery significantly extends our analytical capabilities beyond using experimental data alone.

3. The Possibility of Causal Discovery from Observational Data

Observational data is passively observed, as contrasted with experimental data in which one or more variables is manipulated (often randomly) and the effects on other variables are measured. Observational data is more readily available than experimental data. As observational databases become increasingly available, the opportunities for causal discovery increase.

Traditional statistical thinking says that "correlation does not imply causation." Observational data, however, can be informative regarding which causal relationships do or do not exist. Perhaps the simplest example of such a constraint is the inductive principle that if two variables X and Y are not correlated (or, more generally, are not statistically dependent according to some measure), then X does not cause Y, and Y does not cause X. While this principle can fail, it also can serve as a powerful guide in the search for causal relationships. The story, however, is much richer and more interesting than that simple principle. In particular, a key idea in this book is that among a set of variables, the statistical relationships that are obtained from observational data sometimes can strongly suggest likely causal relationships among a subset of those variables. For example, suppose that in fact X causes Y. By measuring just X and Y, we indeed cannot determine whether X causes Y. So, in that limited sense, correlation does not imply causation. If, however, there is a variable W that is known not to be caused by X or Y, then by examining the statistical independence and dependence relationships among W, X, and Y that are obtained from observational data, it sometimes is possible to infer that X very likely causes Y. Section 7 illustrates how. In some instances, even though we may not be able to induce that X causes Y, we may be able to determine, for example, that Y does not cause X, and thereby constrain the possible causal relationships between X and Y.

In order to show how it is possible to discover causal relationships from observational data, we first need a representation of causality. In the next section, I show how Bayesian networks provide such a representation, which I discuss in some detail.

4. Causal Bayesian Networks

A causal Bayesian network (or *causal network* for short) is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node and a child node, relative to the other nodes in the network. In this section, I discuss nodes, arcs, and their combination within causal networks.

4.1. The Entity Being Modeled

Assume that there is an entity about which we are representing causal relationships. That entity might be a single system or it might be a set of systems. An example of a single system is a manufacturing plant in which we are trying to detect causal relationships in order to improve productivity. An example of a set of entities (units) is a set of patients. There are many medical causal relationships that are of interest, including discovering preventable causes of serious disease. When we model the causal relationships of a set of entities, the component entities may or may not share all of the same causal relationships. The chapters in this book focus primarily on entities that share the same causal relationships. If the entities do not share a common set of causal relationships, then our model of those relationships should be a causal mixture model (see section 4.5.6).

4.2. Nodes

A node represents a variable that characterizes some aspect of the entity being modeled causally. The variable may contain continuous or discrete values. If it contains discrete values, then those values may be ordered or unordered.

The meaning of a node is given by its definition, which for simplicity we will assume is equal to its name; in general, the name is any unique label that identifies the definition. For example, *history of smoking* is the name of node X_1 in figure 1. If a name of a variable is not sufficiently precise, then it may not be possible to know the value to give the variable. Consider again *history of smoking*. This name does not indicate whether we mean smoking cigarettes, cigars, or other materials. The name also does not indicate the amount of smoking required for *history of smoking* to be given the value *yes*, or the time period over which such an amount of smoking must occur. A variable with a name that is insufficiently precise is said to fail the clarity test (Howard and Matheson 1984). But such failure or success is not absolute. Generally, we want all variables in a causal model to have names that pass the clarity test well enough for the purposes to which we plan to apply our model. For the example, the name/definition *patient has smoked one to two packs of cigarettes per day during the past 10 years of his or her lifetime, but did not smoke prior to that time* arguably passes the clarity test well enough for many clinical purposes, even though we have not precisely defined, for example, what we mean by a cigarette.

The value of a variable may represent any aspect of the modeled entity. A value may represent a state of the entity, a change in the state of the entity, or some sequence of changes. The value may or may not contain explicit temporal information. The value of any particular variable may be measured or

missing. If the value of a variable will always be missing, then we say this is a *hidden* or *latent variable;* otherwise, we say it is a *measured variable*. We may have no description or identification of hidden variables in a network other than by their causal and probabilistic relations to observed variables.

4.3. Causal Arcs

Let $\mathbf{X}$ be a subset of the modeled variables. Suppose that there are two different manipulations of the variables in $\mathbf{X}$, called $manipulate_i$ and $manipulate_j$, such that $P(Y \mid manipulate_i(\mathbf{X})) \neq P(Y \mid manipulate_j(\mathbf{X}))$. Now let $\mathbf{X}'$ be a subset of $\mathbf{X}$ such that $P(Y \mid manipulate_i(\mathbf{X}')) \neq P(Y \mid manipulate_j(\mathbf{X}'))$, and for every proper subset $\mathbf{X}''$ of $\mathbf{X}'$ it is the case that either $\mathbf{X}'' = \emptyset$ or $P(Y \mid manipulate_i(\mathbf{X}'')) = P(Y \mid manipulate_j(\mathbf{X}''))$.[2] Note that in general there may be more than one set $\mathbf{X}'$ that satisfies these relationships. We say that each variable X in $\mathbf{X}'$ *causally influences* Y and we place an arc from X to Y in the causal Bayesian network. In words, X is a necessary member of a set of variables whose manipulation is sufficient to change the distribution of Y. (Mackie [1974] contains a detailed discussion of causal sufficiency and necessity.) Note that this characterization of causal influence is relative to the set of modeled variables. The causal influence could be direct or it could be mediated through other measured model variables. Note also that the characterization of causality given here requires that we are able to manipulate just the variables in $\mathbf{X}$ and just the variables in each subset of $\mathbf{X}$, without manipulating (disturbing) other variables. The probabilities used in this analysis may be interpreted from either a frequentist or a Bayesian perspective.

Suppose that there is some variable Z, such that X only causally influences Y through Z. We express this relationship in causal network notation as $X \rightarrow Z \rightarrow Y$. Here X is no longer a direct cause of Y in the network, but rather is an indirect cause. We say that Z is a direct (or immediate) effect of X, and Y is an indirect (nonlocal) effect of X. Similarly, we say that Z is a direct (or immediate) cause of Y and X is an indirect (nonlocal) cause of Y.

If X is a direct cause of Y (relative to a set of modeled variables), that does not mean there are no unmodeled hidden variables (representing hidden processes) that link X to Y. Indeed, there almost always will be. We are not required to represent such hidden variables in a causal graph, however, because their influence is captured by the probability distribution of Y given X. Similarly, if a hidden variable only influences one of the measured variables, as for instance variable Z, we need not represent the hidden variable explicitly, because its influence is represented by the conditional probability distribution of Z given its parents. Thus, a causal network provides a causal abstraction that typically represents certain types of hidden processes only implicitly. If, however, a hidden variable causally influences two or more measured variables, then in general it should be represented in a causal network.

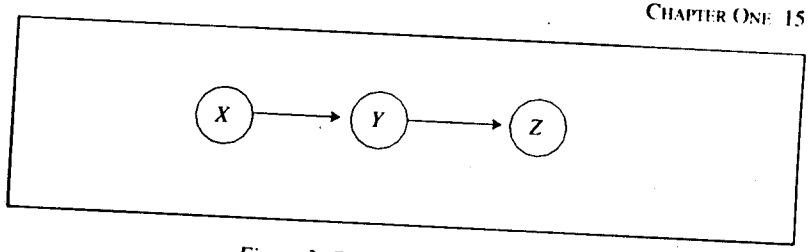

Figure 3. Causal network structure.

4.4. Causal Markov Condition

The causal Markov condition is the Markov condition described in section 1.1 in which arcs are given a causal interpretation. More formally, as adapted from Spirtes, Glymour, and Scheines (1993), the causal Markov condition is as follows:

> Let S be a causal Bayesian network with node set $\mathbf{V}$. Let P be a probability distribution over the nodes in $\mathbf{V}$. The Markov condition is satisfied if and only if for every node X in $\mathbf{V}$ it holds that, according to P, node X is independent of its noneffects (nondescendants) in S given its direct causes (parents) in S.

The intuition underlying the causal version of the Markov condition is as follows. Assume that the structure S of a causal network G is causally valid. A descendant Y of X in S is on a causal path from X. Thus, we would expect there to be the possibility of a probabilistic dependency between X and Y. Now, consider the nondescendants of X; that is, consider all entities represented by the variables in G that are not directly or indirectly caused by X. Let C represent a set of nodes, such that each node in C is a direct and/or indirect cause of one or more parents of X, which we denote as $parents(X)$. Since $parents(X)$ represents all of the direct causes of X, if we do not change the state of these parents, but rather hold just them fixed, we expect that X will be probabilistically independent of each node in C; thus, C will give us no information about the distribution of X. Furthermore, given values for $parents(X)$, we expect that the direct and indirect effects of C (and so on) also are probabilistically independent of X, given just $parents(X)$, unless such an effect happens also to be an effect of X.

The basic intuition underlying the causal Markov condition is that causality is local in time and space. The philosophy literature contains considerable discussion of this issue (Cartwright 1989; Reichenbach 1956; Salmon 1984; Suppes 1970). To illustrate the notion, we now describe two types of locality. According to the causal Markov condition, if we know the local measured causes of X, then nonlocal, indirect causes provide no additional information about the value of X. Consider the causal network in figure 3. Since Y is a local cause of Z, if Y is fixed to some value then changes in the value of X will

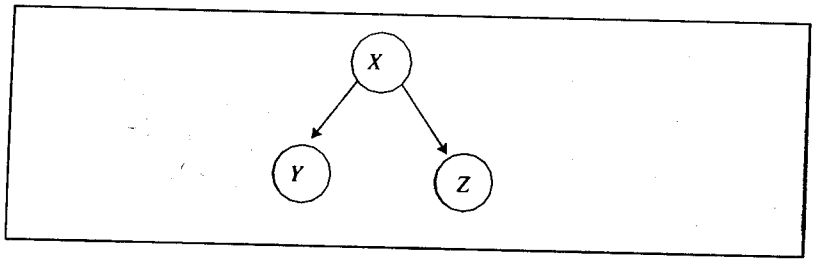

Figure 4. A causal network structure with divergent arcs.

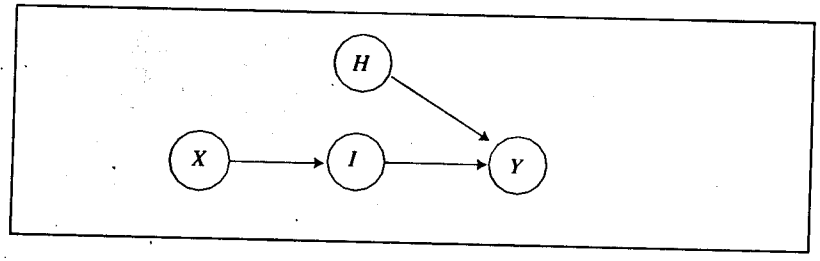

Figure 5. A causal relationship that is probabilistic due to a hidden variable (H).

not causally influence Z. Thus, proximal causes block or screen off more distal, indirect causes.

Consider next the causal network depicted in figure 4. According to the causal Markov condition, if the value of cause variable X is fixed, the value of effect variable Y provides no information about the value of effect variable Z and vice versa. More generally, if there is no directed causal path between two variables, then conditioning on just their common causal ancestors renders the two variables independent.

Causal relationships may be inherently probabilistic (see the end of this section) or probabilistic due to hidden variables. Regarding the latter, consider the causal network depicted in figure 5 in which X, Y, and I are measured variables, and H is a hidden variable. In this situation, X and Y are independent given I. Suppose Y is a deterministic function of H and I, and the probability distribution over H contains no probabilities of 0 or 1. Because of the causal influence of H on Y, the value of Y is not a deterministic function of the value of I; thus, if we are not modeling H explicitly (i.e., it is a hidden causal factor), then we use a probability function to specify a distribution over the values of Y given each value of I.

The causal independence relationships implied by the causal Markov condition should be interpreted relative to (1) the measured and unmeasured variables represented explicitly in the model, and (2) the values of those variables. I discuss both of these provisos next.

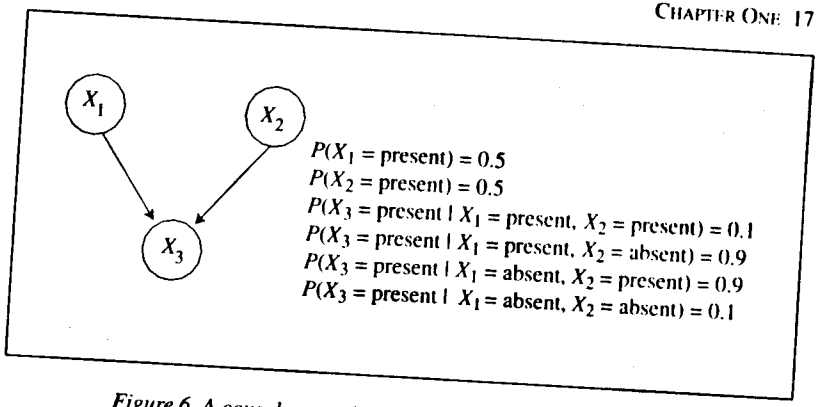

Figure 6. A causal network structure in which X_1 and X_2 taken together (but not alone) have a causal influence on X_3.

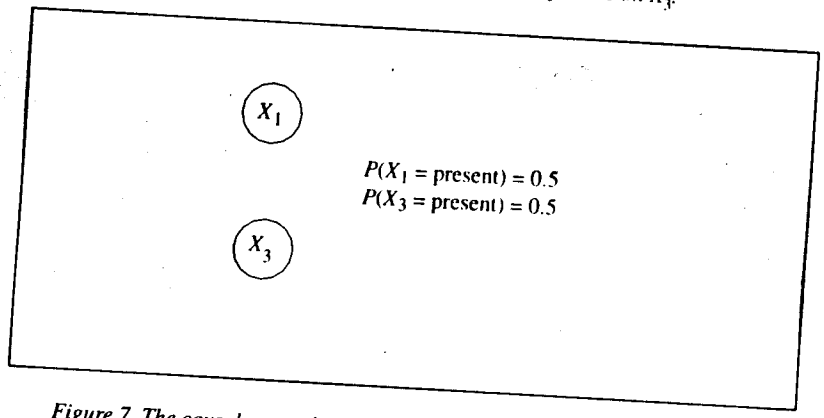

Figure 7. The causal network structure in figure 6, with X_2 marginalized out.

Figure 6 illustrates the first proviso. Suppose there is a causal process that can be represented by the figure. Given the joint probability distribution implied by the probabilities in figure 6, if we consider just variables X_1 and X_3 (i.e., by marginalizing out X_2), we obtain the causal network in figure 7.

As shown in figure 6, when taken together, both X_1 and X_2 causally influence X_3. Figure 7 shows that if we only consider the relationship between X_1 and X_3, then X_1 does not (by itself) have a causal influence on X_3. Consider an RCE that involves just the two variables X_1 and X_3. If we manipulate X_1 and measure X_3, the RCE will show no causal influence of X_1 on X_3. This simple example illustrates yet another way in which causality must be interpreted relative to the set of variables that are in the model. The absence of a statistical dependency between two variables does not mean they are causally unrelated to one another when unmeasured variables are considered.

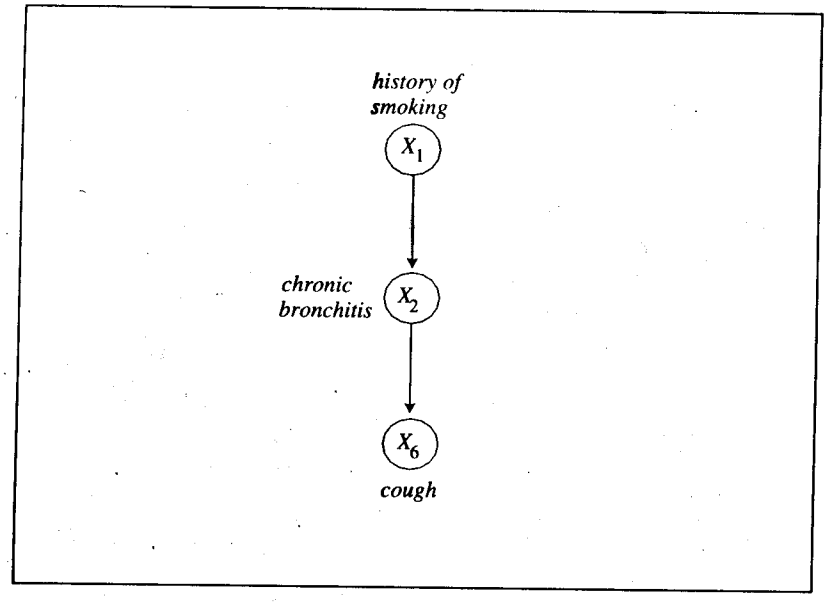

Figure 8. A simple three-node causal network structure.

Note that this example contains a special type of distribution that violates the faithfulness assumption discussed in section 6.1; for most distributions, marginalizing out X_2 would not lead to a situation in which X_1 has no causal influence on X_3. Still, the distribution of X_3 given X_1 may be significantly different when X_2 is in the model than when it is not. For example, there exist distributions for which X_1 having the value present, when considered alone, makes X_3 highly likely to have the value present; however, when a variable X_2 has the value present, then X_1 having the value present makes X_3 highly likely to have the value absent. *The general point, then, is that the causal relationships in a model should be interpreted relative to the variables in that model* (Aliferis and Cooper 1998).

The network structure depicted in figure 8 illustrates the second proviso previously mentioned, which involves the values of variables. Suppose that the three variables can take on a value from the following respective companion sets:

history of smoking: {none, moderate, severe}
chronic bronchitis: {absent, moderate, severe}
cough: {absent, present}

It could well be that given this value representation, then the Markov condition applied to the above network would imply the correct independence relationships: namely, the value of *cough* is independent of the value of *history*

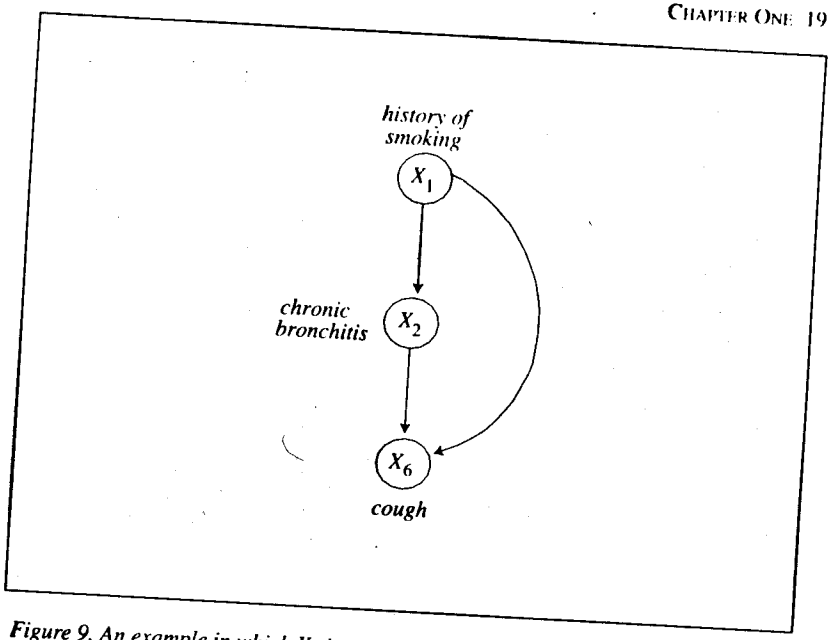

Figure 9. An example in which X_6 is not independent of X_1 given X_2 because the possible values of X_2 are too coarse to sustain that independence.

of smoking, given that we know the value of *chronic bronchitis*.

Consider, however, the following representation of the values, which for chronic bronchitis are coarser than before:

history of smoking: {none, moderate, severe}
chronic bronchitis: {absent, present}
cough: {absent, present}

Suppose we condition on *chronic bronchitis* having the value present. It may be the case that a severe degree of smoking suggests (with high likelihood) a severe degree of *chronic bronchitis*, which in turn suggests a high likelihood of a *cough* being present. Similarly, a moderate degree of *smoking* suggests a moderate degree of *chronic bronchitis*, which suggests an intermediate likelihood of a *cough*. In this situation, *cough* is not independent of *history of smoking*, given that all we know is that *chronic bronchitis* has the value present; the causal network in figure 9 expresses the independence relationships that exist (namely, none) given the coarser variable-value representation being used for *chronic bronchitis*.

Is the arc from *history of smoking* to *cough* causal in figure 9? In a sense, yes. If we maintain (i.e., fix) *chronic bronchitis* to the value present and we manipulate *history of smoking* between values of none, moderate, and severe, then we would expect (given this story) that the likelihood of *cough* will vary

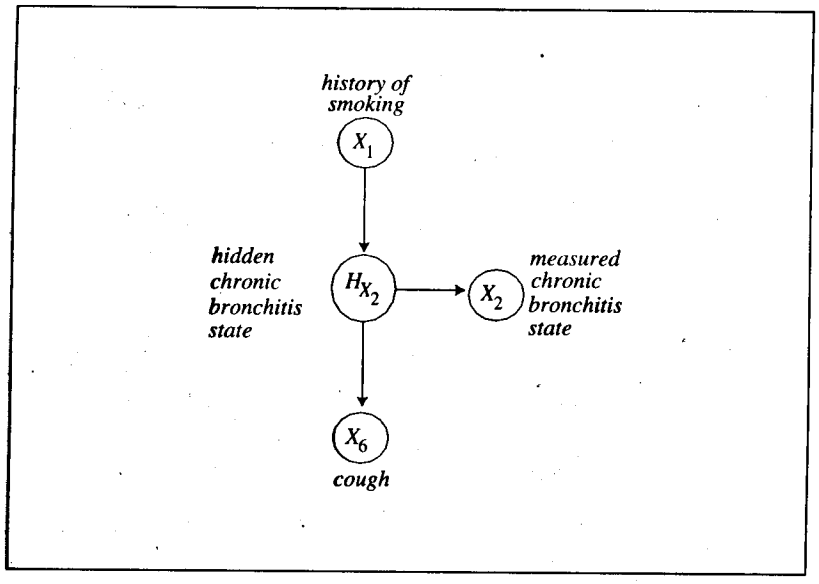

Figure 10. An alternative representation of the causal network structure shown in figure 9.

with the value of *history of smoking*. The reason is that even though we maintain *chronic bronchitis* to just the value present, it can (and probably will) still vary between being moderate and severe. That is, in maintaining *chronic bronchitis* to be present, we only ensure that the value present holds, we do not ensure that any more specific value of the variable holds or does not hold.

Figure 9 illustrates only one specific situation in which the values of network variables are important; there are other more complex situations. The general point, however, is that there is no fixed, directed acyclic graph structure for which the causal Markov condition will be valid, if we allow the value sets of the variables to vary. In other words, the interpretation of the arcs in a causal network is relative to the sets of values associated with the nodes in the network. Thus, those sets of values should in essence be considered a part of the causal network structure. When we "read" causal relationships from the directed acyclic graph, our interpretation of the arcs in that graph should be relative to the value representation of the nodes in the graph.

An alternative representation of the causal network in figure 9 is shown in figure 10, where H_{X_2} has the value set {absent, moderate, severe} and X_2 has the value set {absent, present}. Variable X_2 is a deterministic function of H_{X_2}, but not vice versa. The basic idea underlying the representation in figure 10 is that measured variables can be abstractions of unmeasured variables (Aliferis and Cooper 1998). According to the causal Markov condition, conditioning on X_2 does not render X_1 and X_6 independent.

I close this section with a brief discussion of one additional issue regarding the causal Markov condition. Current quantum theory and experimental data provide support for the notion that at the quantum level, causality is not temporally and spatially local (Herbert 1985). Thus, at that level, there may be no local variables that render a variable X independent of other variables. Such a situation would not imply that the Markov condition is violated, but rather that quantum events are less independent than we intuitively expect them to be. The utility of the Markov condition might thereby be diminished in the discovery of causal relationships at the quantum level. I note also that quantum events can be coupled tightly to macroscopic events (Herbert 1985; Zurek 1991); thus, it may be that the Markov condition would hold for macroscopic systems less often than we intuitively expect. Nonetheless, our common experience suggests that local conditioning as given by the causal Markov condition often does hold in the macroscopic world, at least approximately. The methods described in this book employ the causal Markov condition as a reasonable working assumption. It is an open problem to investigate the frequency with which we expect the condition to hold in the macroscopic world and yet it does not.

4.5. Causal Relationships

This section introduces several important concepts regarding causal relationships, including temporal representation, transitivity, multivariate causes, confounding, selection bias, and compound relationships.

4.5.1. Temporal Representation

Often the temporal relationship between a cause and an effect is left implicit in a causal network. Consider the nodes *history of smoking* and *chronic bronchitis* in figure 8. For simplicity, assume that we are using the causal network to predict *chronic bronchitis* given a *history of smoking*. Suppose that $P(\textit{chronic bronchitis} = \text{present} \mid \textit{history of smoking} = \text{severe}) = 0.3$. What does this probability mean exactly?

We view *history of smoking* as taking on the value severe some time in the past. But when in the past? And what was the value of *history of smoking* from which we manipulated it to the value severe? These details are unspecified here. A problem may arise if *history of smoking* has a modeled cause C. In that case, the value of C may change the distribution over the possible ways that *history of smoking* takes on the value severe. Thus, depending on the value of C, we may have a different distribution for *chronic bronchitis* conditioned on *history of smoking*. The problem is one of an insufficient representation of the values of *history of smoking*. It is the same type of phe-

nomenon that I discussed in association with figures 9 and 10, except specialized here to be the temporal dimension of the value set. A solution is to represent the temporal dimension of the values of *history of smoking* in finer detail, so that conditioning on *history of smoking* makes *chronic bronchitis* independent of ancestors of *history of smoking*.

Researchers have begun to develop temporal Bayesian networks that permit detailed explicit modeling of temporal causal relationships among variables (Aliferis 1998; Aliferis and Cooper 1995; Aliferis and Cooper 1998; Berzuini, Bellazi, Quaglini, and Spieglehalter 1992; Dagum and Galper 1993; Provan and Clarke 1993). These representations provide the basis for future learning algorithms that can induce detailed causal temporal patterns from data.

4.5.2. Transitivity

Causality is not necessarily transitive. Although X may cause I and I may cause Y, it could be that X does not cause Y. Violations of transitivity require special probability distributions in order to exhibit nontransitivity. Consider the following example for which X is a binary variable that takes the values x_1 and x_2, Y is a binary variable that takes the values y_1 and y_2, and variable I can take on any one of the values i_1, i_2, i_3, or i_4.

$P(X = x_1) = 3/4$
$P(I = i_1 \mid X = x_1) = 2/9 \quad P(I = i_1 \mid X = x_2) = 2/9$
$P(I = i_2 \mid X = x_1) = 4/9 \quad P(I = i_2 \mid X = x_2) = 1/9$
$P(I = i_3 \mid X = x_1) = 1/9 \quad P(I = i_3 \mid X = x_2) = 4/9$
$P(I = i_4 \mid X = x_1) = 2/9 \quad P(I = i_4 \mid X = x_2) = 2/9$
$P(Y = y_1 \mid I = i_1) = 2/3$
$P(Y = y_1 \mid I = i_2) = 1/2$
$P(Y = y_1 \mid I = i_3) = 1/2$
$P(Y = y_1 \mid I = i_4) = 1/3$

In the distribution defined by this causal Bayesian network, X and I are dependent, and I and Y are dependent, yet X and Y are independent because $P(Y = y_1 \mid X = x_1) = P(Y = y_1 \mid X = x_2) = 1/2$.

4.5.3. Multivariate Causes and Explaining Away

When a node X has more than one parent node, those parent nodes often have a characteristic pattern of being dependent conditioned on X. To illustrate, consider the causal network in figure 11, in which we assume X_2 and X_3 are marginally independent and are both causes of X_4. While X_2 and X_3 are marginally independent when we do not condition on X_4, they often are dependent when we condition on X_4. Since *chronic bronchitis* and *lung cancer* are causally independent (as assumed here), it is appropriate that they are marginally independent. Consider conditioning on *fatigue*. If *fatigue* is present, then the presence of *chronic bronchitis* and *lung cancer* each become

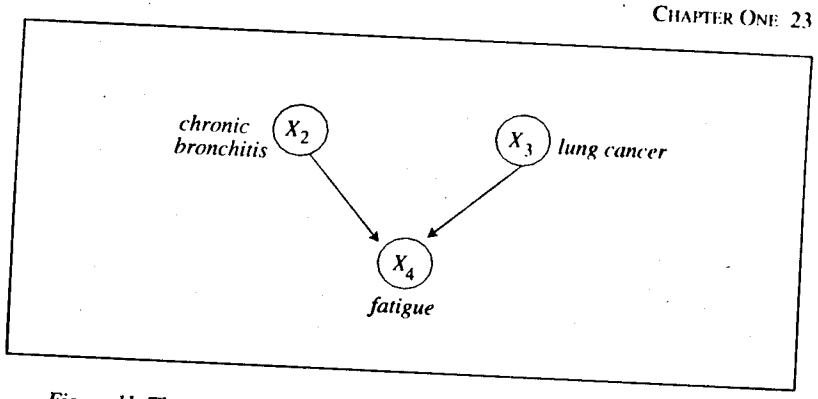

Figure 11. The structure of a causal network that illustrates explaining away.

more likely. This likelihood of *lung cancer* will decrease, however, if we learn that *chronic bronchitis* is present, because *chronic bronchitis* provides an explanation for the *fatigue* (i.e., chronic bronchitis "explains away" fatigue). Thus, conditioned on *fatigue*, the events *chronic bronchitis* and *lung cancer* are not probabilistically independent.

The appearance of "explaining away" is one example of how parent nodes can be dependent given that we condition on the child node (Wellman and Henrion 1993). Other types of conditional dependency can exist. Although in causal Bayesian networks there is not a requirement that parents be dependent conditioned on some value of their child, often they are (Pearl 1988); for example, in linear models and in noisy-or-gate models, they must be.

4.5.4. Confounding

If two variables are probabilistically dependent due (at least in part) to one or more shared causes (either direct or indirect), then the two variables are said to be confounded and their common causes are called the *confounders*. The confounders can be either measured or unmeasured variables. The causal network structure in figure 12, which is taken from figure 1, shows *history of smoking* as a confounder of *chronic bronchitis* and *lung cancer*.

Confounding is important, because when two variables X and Y are statistically dependent, often the most likely possibilities are that (1) one variable causally influences the other, (2) the two variables are confounded, or (3) both 1 and 2 hold. If the confounders are measured, we can condition on them and remove the statistical dependency between X and Y that is due to confounding. If the confounders are not measured, then we need other methods for detecting or eliminating them.

An RCE is one way to eliminate a confounder. For example, if we manipulate *chronic bronchitis*, then we break the arc from *history of smoking* into *chronic bronchitis*, because our manipulation means that *history of smoking*

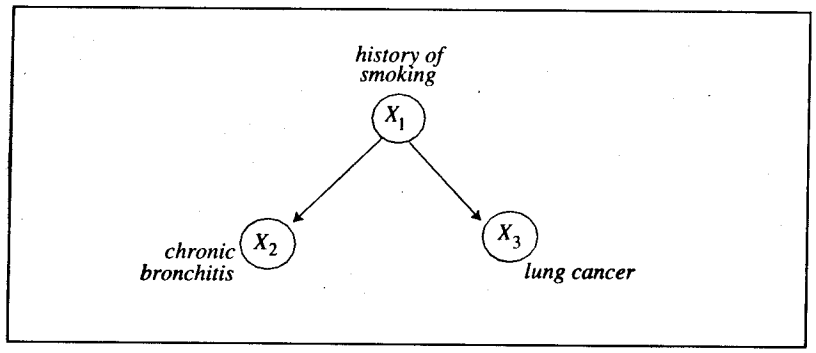

Figure 12. History of smoking is the confounder of chronic bronchitis and lung cancer.

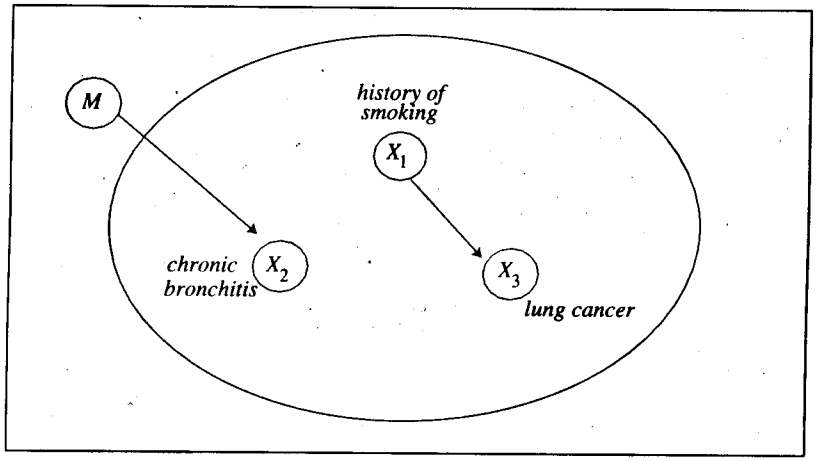

Figure 13. An RCE is one way to eliminate a confounder.

no longer has a causal influence on *chronic bronchitis*. Thus, we would obtain the causal network shown in figure 13. Data generated by a process that is represented by figure 13 is expected to support *chronic bronchitis* as being independent of *lung cancer*, which would support that *chronic bronchitis* is not a cause of *lung cancer*.

Section 7 contains a discussion of assumptions under which observational data is sufficient to determine that two variables are statistically dependent due to confounding.

4.5.5. Selection Bias

If V' denotes an arbitrary instantiation of all the variables in V, then we want

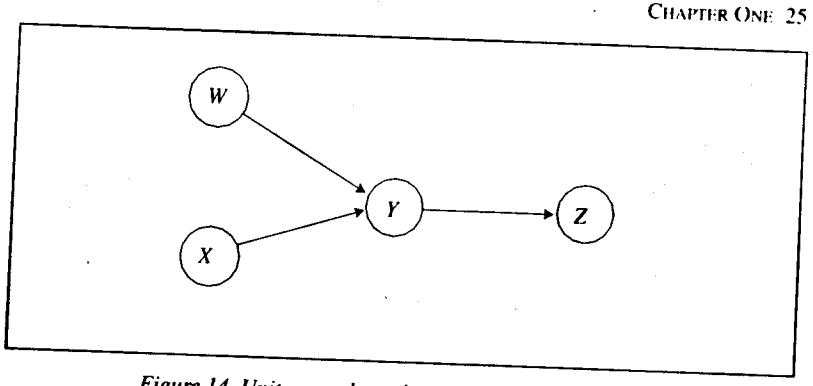

Figure 14. Units may share this causal network structure but differ in the distribution of Z given Y.

that V' is sampled for inclusion in D with probability $Pr(V' \mid G)$, where G is the causal Bayesian network to be discovered, which represents the data generating causal process. If selection bias exists, then V' is sampled with a probability other than $Pr(V' \mid G)$.

Suppose that an individual with only a fever (X) is likely to stay home and take aspirin. Similarly, a person with only abdominal pain (Y) is likely to stay home and take an over-the-counter medication for relief. Suppose, however, that an individual with both fever and abdominal pain is likely to be concerned about the possibility of a serious illness, and therefore, is prone to go to his or her local emergency room, where we have been collecting our data. In this situation, X and Y may be dependent, due to selection bias, even though X does not causally influence Y, Y does not causally influence X, and X and Y have no common confounder. Such bias can persist, regardless of how large the sample size. Selection bias can be avoided in RCEs by measuring the outcomes that follow in time for each unit in the experiment (e.g., the outcomes for each patient in a clinical trial). The presence or absence of selection bias sometimes can be inferred from observational data (Cooper 1995b). Chapter 6 provides a detailed handling of selection bias when using constraint-based methods for causal discovery from observational data.

4.5.6. Causal Mixtures

Recall that in section 4.1 I stated that usually the units under study are assumed to share a common set of causal relationships, both in terms of causal structure and the parameterization of that structure. It could be, however, that the members of a given population of units do not all share the same causal relationships. For example, all units may share the same causal network structure (figure 14), but one subpopulation of units may have a different distribution between Y and Z than does the remaining subpopulation.

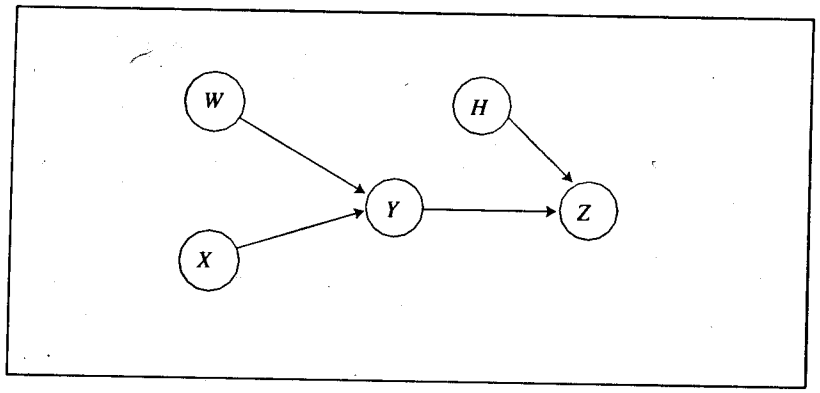

Figure 15. A hidden variable H can represent a mixture of subpopulations that differ according to the parameterization of Z given its cause Y.

We can represent such a mixture by using a hidden binary variable H, which is included in the causal network structure shown in figure 15. More generally, hidden variables seem adequate to represent mixtures that involve just the parameters relative to a common structure (rather than mixtures of causal network structures).

The structure depicted in figure 15 represents at least two possible situations. In one, which I just discussed, there is a mixture of units represented by H, with each subpopulation having a common distribution (possibly deterministic) of Z given Y. In the other situation, all the population units are homogeneous and each unit has the same distribution of Z given Y; the variable H serves only to represent the inherent uncertainty expressed by that distribution.

In an extreme case, one subset of the population might have the causal network structure shown in figure 16a, and the remaining subset might have the network shown in figure 16b. Causal structure includes both the arcs among nodes and the value range of each node. In general, admitting mixtures weakens our ability to learn causal relationships from observational data. The discovery of mixtures of causal structures is a challenging, largely open problem.

One approach to admitting mixtures is to have a graphical language that expresses them as members of an equivalence class of causal networks that are statistically indistinguishable. Another approach would be to search over subsets of the cases to locate for each subset a causal network (or set of networks) that is most likely given the subset. A Bayesian version of this approach would require specifying a prior probability over the various ways of forming subsets of the cases in the database. This approach makes possible

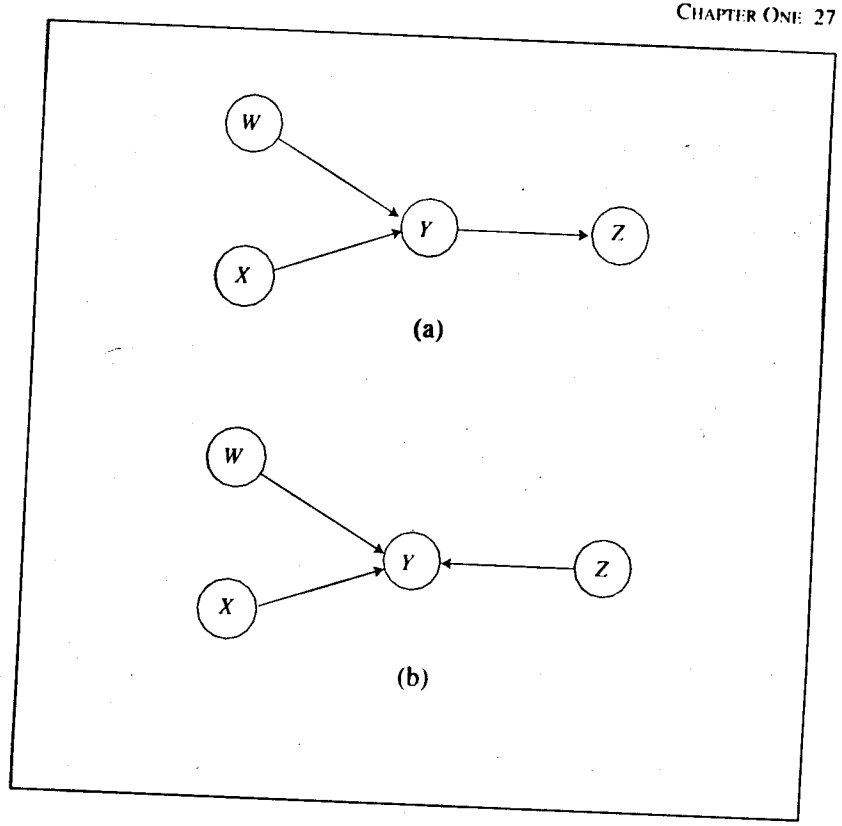

Figure 16 a and b. The causal network structures for two subpopulations.

the scoring of causal relationships for subsets of the population; from such analyses we could derive the posterior probabilities of causal relationships for the population as a whole. A challenge in applying this Bayesian approach is to find ways to make it computationally tractable.

4.5.7. Compound Relationships

The statistical dependency between two measured variables X and Y may be due to a complex combination of the following mechanisms: (1) direct causality, (2) indirect causality, (3) confounding, and (4) selection bias. The Bayesian network structure in figure 17 illustrates one such possibility, which involves all four mechanisms. The shaded node S represents a special instantiated variable that indicates selection of a unit for observation based on the values of the parents of S, namely X and Y in the current example. There are of course many other combinations of the four mechanisms. For

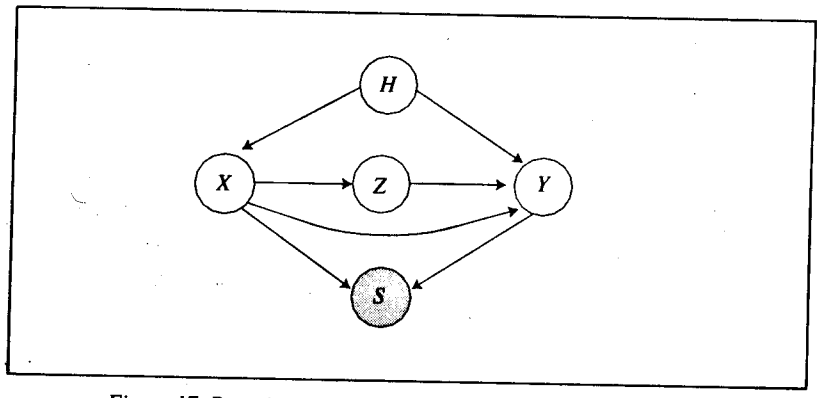

Figure 17. Bayesian network structure involving direct causality, indirect causality, confounding, and selection bias.

example, suppose we limit consideration to just the five nodes in the figure 17 network. Consider every directed acyclic graph on those nodes for which X and Y are d-connected (conditioned on no other variables). Each such graph represents a set of mechanisms which can render X and Y statistically dependent.

5. Causal Discovery

Up to this point, I have described the causal Bayesian network as a representation of causal relationships, and I have discussed some of its properties. In the remainder of this chapter, I discuss two approaches for discovering causal Bayesian networks from observational data. One approach, which is described in sections 6 and 7, uses tests of conditional independence and dependence among subsets of model variables to constrain the causal relationships among the model variables. Another approach, which is described in section 8, computes the probability that causal relationships exist among the variables. Section 9 introduces several basic search algorithms that have been used with these two fundamental approaches to causal discovery. These sections are not intended to provide a comprehensive overview of research on causal discovery methods (see, for example, Angrist, Imbens, and Rubin 1996; Balke and Pearl 1994; Bollen 1989; Bowden and Turkington 1984; Heise 1975; Manski 1995; Meyer 1995; Pearl 1996; Pratt and Schlaifer 1988; Robins 1986, 1989; Rubin 1974; Simon 1953; and Wright 1921 for samples of relevant prior work), but rather to introduce basic concepts for the graphical causal discovery methods that are emphasized in this book.

An assumption that typically is made with the causal discovery approaches described in this book is that the samples (i.e., cases, records, instances, etc.) in the observational database are *independently sampled* and *identically distributed*. Independent sampling means that given a causal Bayesian network model, the probability of one sample is independent of any other samples that have been obtained. In theory, a lack of independence among cases could be modeled using hidden variables (see Spirtes, Glymour, and Scheines 1993, section 9.4), but this makes learning causal relationships more difficult and is seldom done. Samples are identically distributed if the probability of seeing a given case at one point in time is the same as seeing that case at another point in time; that is, the joint probability distribution defined by the causal Bayesian network is time invariant. The assumption of identically distributed cases can be relaxed by using temporal causal Bayesian networks that explicitly allow the representation of changes in distributions over time. While progress is being made on the representation of temporal Bayesian networks (see section 4.5.1), relatively little research has been done yet in learning such representations from time series data.

If a set of causal structures can equally account for the same observational data, then no observational data can distinguish among them. This fundamental concept is a type of *statistical indistinguishability*. If no members of a set of causal structures are statistically indistinguishable, then they are called *statistically distinguishable*, or equivalently, *statistically identifiable*. Different types of statistical indistinguishability are established based on different meanings of the phrase "can equally account for" in the preceding sentence. Returning to a previous simple example, consider the causal networks $X \rightarrow Y$ and $X \leftarrow Y$. If we do not restrict the distributions considered, then both networks can represent any joint probability distribution on the two variables, and thus, they each can equally well account for the same observational data. Therefore, we would say that the two networks are statistically indistinguishable given only observational data. By contrast, the causal network X no_arc Y is in general statistically distinguishable from $X \rightarrow Y$, because with observational data X and Y will be statistically independent in the former network, whereas in the latter network they will not. For a more detailed discussion of statistical distinguishability in causal discovery, see chapter 4 of Spirtes, Glymour, and Scheines (1993). A key theme underlying the topics in this book is that there are interesting classes of causal networks that under assumptions are statistically distinguishable, based on observational data. Section 6 introduces one such set of assumptions. Section 10 provides some support for these assumptions, although further evaluation is an important open problem.

6. Assumptions for Constraint-Based Causal Discovery

In this section I discuss the assumptions typically made in constraint-based methods for discovering causal knowledge from observational data. Almost always the methods assume that the causal processes generating the data can be modeled as a Bayesian network; in this chapter, for brevity, I sometimes state that the Bayesian network itself generated the data. Since, as described in section 1.1, the Markov condition is inherent in the Bayesian network representation, the discovery methods assume the causal Markov condition. Recently, researchers have begun to extend the Bayesian network representation of causal relationships; one extension is the representation of causal feedback cycles with directed *cyclic* graphs (see chapter 7).

Constraint-based causal discovery involves a two-step procedure in which (1) statistical tests are used to establish conditional dependence and independence relationships among the variables in a model, and (2) those relationships are used to constrain the types of causal relationships that exist among the model variables.

The remainder of section 6 summarizes and illustrates typical assumptions that have been used in applying constraint-based causal discovery methods. Chapters 2, 3, 5, and 6 describe these assumptions and their use in additional detail. Chapters 8, 9, 10, and 11 contain arguments for and against some of these assumptions holding for causal discovery in the real world.

6.1. Causal Faithfulness Assumption

Let G be a causal Bayesian network, V be the nodes in G, S be the network structure of G, and P be the joint probability distribution generated by G. The *causal faithfulness assumption* is as follows:

> For all disjoint sets A, B, and C in V, if in S we have that A is not d-separated from B given C, then in P we have that A and B are conditionally dependent given C, where A and B are not empty but C may be.

The causal faithfulness assumption says that the only way variables will be probabilistically independent is if their independence is due to the Markov condition, or equivalently, to the d-separation condition. In other words, if variables are d-connected (i.e., not d-separated) in G then they are dependent in P. Thus, the network structure S reveals all the independence relationships among all the variables in V relative to the underlying distribution P. Note that P must be estimated from data D; we generally do not know P exactly.

For example, in the network in figure 1, consider just nodes X_1 and X_2, which are d-connected by the arc between them. The faithfulness assumption

would be violated if the probabilities given in figure 1 were changed so that X_1 and X_2 are marginally independent, but the arc from X_1 to X_2 remained. In the actual joint probability distribution that follows from the probabilities given in figure 1, the faithfulness assumption is not violated.

The Markov condition relates causal structure to probabilistic independence and the faithfulness assumption relates causal structure to probabilistic dependence. Together, they provide a highly informative mapping between the independence and dependence relationships of model variables as given by the their probability distribution and the d-separation/d-connection relationships of the corresponding nodes in a Bayesian network structure. Thus, we can use statistically inferred independence and dependence relationships (see section 7) to constrain the structure of the Bayesian network that is generating the data.

The following result regarding the faithfulness assumption has been proved for discrete (Meek 1995b) and for multivariate Gaussian (Spirtes, Glymour, and Scheines 1993) Bayesian networks. Consider any *smooth* distribution[3] Q over the possible parameters in a Bayesian network. The parameters are just the probabilities represented in the network. Now consider drawing a particular set of parameters from distribution Q. The results in Meek (1995b) and Spirtes, Glymour, and Scheines (1993) show that the probability of drawing a distribution that is not faithful is Lebesgue measure zero. These results do not mean that drawing such a distribution is impossible, but rather, under the assumption of a smooth distribution, such an outcome is exceedingly unlikely.

Most current constraint-based causal discovery methods, including those described in this book, are based on the faithfulness assumption. Alternative assumptions, such as the minimality assumption, also have been considered (Yao and Tritchler 1996; Spirtes, Glymour, and Scheines 1993). These alternative approaches are not, however, discussed further in this chapter.

While the faithfulness assumption is plausible in many circumstances, there are circumstances in which it is invalid. In the remainder of this section I outline some basic reasons that the faithfulness assumption can fail.

Deterministic relationships can interfere with causal discovery from observational data. Consider the following causal Bayesian network structure

$X \rightarrow Y \rightarrow Z$

for which all three variables are binary and

$P(X = \text{yes}) = p$,
$P(Y = \text{yes} \mid X = \text{yes}) = q$,
$P(Y = \text{no} \mid X = \text{no}) = q$,
$P(Z = \text{yes} \mid Y = \text{yes}) = q$, and
$P(Z = \text{no} \mid Y = \text{no}) = q$.

For the moment, assume $p = 1$ and $q = 1$. Thus, the three variables are de-

terministically related, and indeed, they always have the same value *yes*. Since there is no variation in the values of variables, we cannot determine from observational data what would happen if variation (in the form of manipulation) were to take place.

Consider next that $p = 0.5$ and again $q = 1$. The Markov condition applied to the network structure of the example does not imply (through d-separation) that X is independent of Y given Z, and thus, by the faithfulness assumption such independence should not hold. But, for the distribution defined, X is independent of Y given Z. Knowing the value of Z tells us the value of X exactly, and therefore, conditioning on Y makes no difference. In the example, the faithfulness assumption is valid for any value of q that is not equal to 0, 0.5, or 1. Practically, however, as q gets close to 0, 0.5, or 1, the usefulness of the assumption being technically valid begins to decrease, because with finite data samples the variables will appear to be deterministically related (for $q = 0$ or $q = 1$) or independent (for $q = 0.5$), and thus the faithfulness assumption will appear to be violated.

Violation of the faithfulness assumption does not, however, require the presence of deterministic relationships, as shown previously by the example given in figures 6 and 7. Here each of X_1 and X_2 considered alone is marginally independent of X_3. When X_1 and X_2 are taken together, however, there is a dependency between them and X_3. Other nondeterministic distributions that violate the faithfulness assumption are described in Spirtes, Glymour, and Scheines (1993), including distributions based on special cases of Simpson's paradox, which I briefly describe next.

Qualitatively, Simpson's original example (Simpson 1951) is as follows. Consider a population of people. Among males in the population there is a positive statistical association between receiving a particular treatment and surviving. Similarly, among females there is a positive statistical association between receiving a particular treatment and surviving. However, when considering the population as a whole (both males and females), there is no statistical association between the treatment and survival. Suppose, in reality, that treatment causally influences survival and that gender confounds treatment and survival. Then, the example distribution violates the faithfulness condition, relative to a Bayesian causal network that represents the causal reality. Two additional points are worth noting here. First, not surprisingly, very special distributions are required to exhibit Simpson's paradox. Second, in the example just given, the statistical associations among the variables representing *gender*, *treatment*, and *survival* are based on observational data. The example does not indicate that the paradox would persist under manipulation in an RCE, and indeed, it would not.

Goal-oriented systems, both animate and inanimate, provide another general class of situations in which violations of the faithfulness conditions may tend to occur. Consider a generic clinical situation that is modeled in figure 18. As-

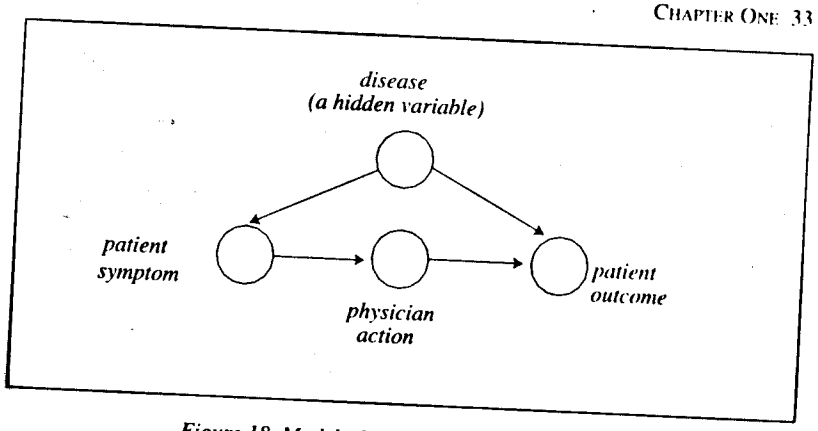

Figure 18. Model of a general clinical situation.

sume that the patient outcome is *near-term mortality*. A disease causes a patient symptom, which the physician observes and which influences the physician's action to obtain a cure. Suppose that if the disease is severe enough to cause a patient's death, then the symptom appears and the physician initiates a treatment action. Assume further that the treatment always prevents near-term mortality. Based on information about their patients, physicians take actions that maintain patient outcomes to be the preferred state of surviving. This physician goal-oriented behavior leads to near-term survival of all patients; thus, the physician action variable and the patient outcome variable will be independent, which violates the faithfulness assumption. This example does contain probabilities of one and zero. More significantly, however, is that the example illustrates a system (here a physician) whose goal (implicitly) is to violate distributions that would follow from the faithfulness assumption. Although the goal may not be completely achieved in practice, the induced distribution may be close enough to being unfaithful that it makes causal discovery difficult; this point is supported by an experiment described in chapter 15 that involves the photosynthetic rate and internal CO_2 concentration in plant leaves. The faithfulness assumption would more plausibly hold for systems (or subsystems) that are not goal-oriented.

In the context of all possible distributions on a set of variables, there are relatively few unfaithful distributions. Thus, a violation of the faithfulness assumption is not likely unless we have reason to believe that such special distributions are present (Meek 1995b; Spirtes, Glymour, and Scheines 1993). This section has described several cases in which the likelihood of occurrence of such special distributions is heightened. The existence of such unfaithful distributions can lead to errors by causal discovery methods that assume faithfulness.

6.2. The Assumption of Valid Statistical Testing

In attempting to discover causal relationships from observational data, we do not have a probability distribution for the underlying causal process that is generating the data, we just have the data. Thus, we need some way of linking inference of independence and dependence relationships from data to the underlying probability distribution on which the Markov condition and faithfulness assumption are based. The following assumption regarding valid statistical testing does just that:

> Consider the sets of variables A, B, and C in V. If in the underlying distribution P we have that A and B are conditionally dependent given C, then A and B are conditionally dependent given C according to test T applied to the data in D. Similarly, if in P we have that A and B are conditionally independent given C, then A and B are conditionally independent given C according to test T applied to the data in D.

We are assuming that test T can be used to uncover the probabilistic dependence and independence relationships among the measured variables, as given by P. Note that T implicitly includes the value of any statistical significance threshold (e.g., an alpha level) that is required in applying the test.

The smaller the number of cases in D, the more skeptical we should be of whether statistical testing is valid. When using classical statistical tests of independence, such as the chi-square test, it is not clear, even for a large database, precisely which value to use as a statistical threshold. The Bayesian causal discovery methods (see section 8) avoid categorical tests of independence and dependence, and instead use a continuous measure for scoring networks that inherently encodes the uncertainty of small data samples. Closely related methods, based on using minimum description length scores or entropic measures, also avoid categorical tests (Bouckaert 1995, Herskovits and Cooper 1991, Lam and Bacchus 1994, Wedelin 1993).

As mentioned in section 6.1, there is an interplay between the faithfulness assumption and the assumption of valid statistical testing. As distributions approach being unfaithful, we require more data in order for statistical tests to reliably detect dependence among network variables. Chapters 8–11 provide further discussion of this issue.

6.3. Missing Data

Constraint-based causal discovery programs may make one of several assumptions about how to handle missing data. I discuss several possible approaches in this section.

Consider a database D that contains records (e.g., patient cases). Often there is missing data in D; that is, each variable is not measured for each record. One approach to dealing with missing data is to remove all records

from D in which any of these variables has a missing value. The problem with this approach is two-fold. First, we may end up with a very small database (possibly even zero records) for learning. Second, the data may not be missing randomly, in which case the distribution among the complete records may not accurately reflect the distribution in the unselected population of interest. Although sometimes we may be able to detect such selection bias, it can interfere with uncovering causal relationships that exist.

Another solution to the problem of missing data is to assign the value *missing* to a variable in a record for which that variable was not measured. This approach may, however, lead to the loss of independence relationships that otherwise would hold were all the data measured. Consider the causal network structure $X \rightarrow Y \rightarrow Z$ as a valid representation of the causal relationships among the three variables when each is measured. Conditioned on Y having the explicit value *missing*, the value of X may provide some information about the value of Z, and thus, X and Z may test as being dependent conditioned on Y.

A third solution to the problem is to *fill in* each missing value of each variable with some admissible value for the variable. There are numerous methods for assigning missing values (Little and Rubin 1987). Hopefully, of course, the substituted values correspond closely to the actual, underlying values, but in general there is no guarantee that this will be the case.

6.4. Types of Variables and Distributions

In principle, constraint-based discovery methods apply when there are continuous variables, or even a mixture of continuous and discrete variables, as long as there are reliable statistical tests of independence and dependence. Numerous statistical tests of independence exist for discrete variables. For multivariate Gaussian Bayesian networks, tests also exist (Spirtes, Glymour, and Scheines 1993). Developing statistical tests that apply to a wide variety of distributions on continuous variables (or mixed continuous and discrete variables) is an open problem.

7. Constraint-based Methods for Evaluating Causal Bayesian Networks

In this section I first provide a simple illustration of the constraint-based causal discovery method. Then I briefly discuss two representative algorithms for constraint-based causal discovery.

Figure 19 extends figure 1 by adding a causal arc from a node representing

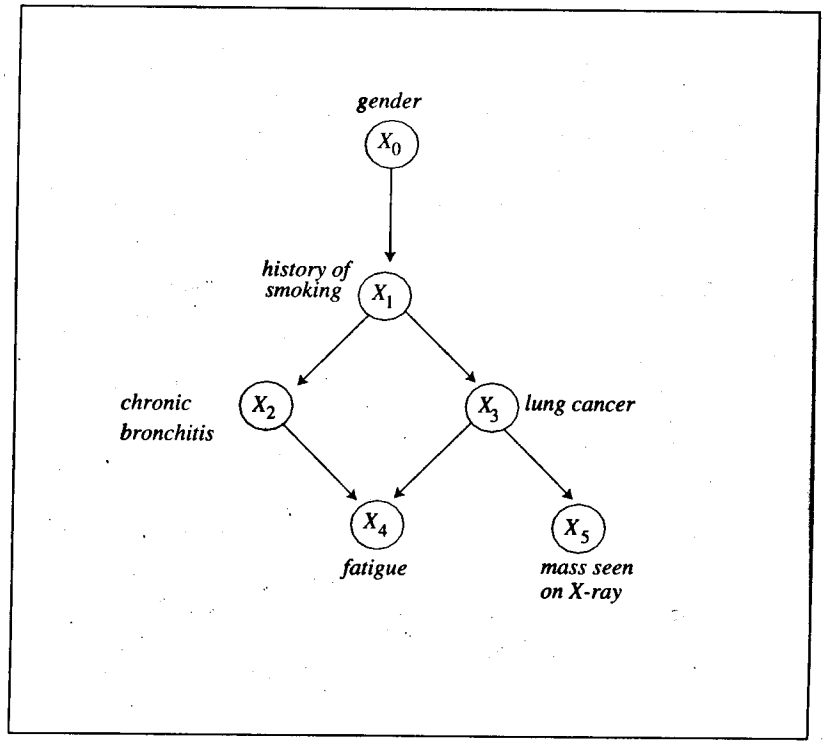

Figure 19. An extension of figure 1 with the node gender added.

gender (X_0) to a node representing *history of smoking* (X_1). Does smoking cause lung cancer? For the purpose of this simple example, suppose we know, or are willing to assume, that gender causally influences whether an individual smokes. Consider the four causal Bayesian networks in table 2 as representing among the three variables the set of causal relationships that we believe are tenable. Suppose that smoking actually does cause lung cancer. Given the assumptions in section 6, we can infer the *d*-separation conditions in row 2 of table 2, which are unique relative to the *d*-separation conditions of the other three possible causal hypotheses being considered. Thus, under the assumptions made, we can conclude that *history of smoking* is causally influencing whether a patient gets *lung cancer*. We can quantify that causal relationship by using the data to estimate $P(\textit{lung cancer} \mid \textit{history of smoking})$. In general, if we know that the set of variables in T causes variable Y without confounding or selection bias, then we are warranted in equating $P(Y \mid \textit{manipulate}(T))$ with $P(Y \mid T)$, which we can estimate from available observational data.

Row no.	Structure of a causal network C	d-separation conditions as inferred by statistical tests	
		$DS(X_0, X_3)$	$DS(X_0, X_3 \mid X_1)$
1	$X_0 \quad X_1 \quad X_3$	+	+
2	$X_0 \rightarrow X_1 \rightarrow X_3$		+
3	$X_0 \rightarrow X_1 \leftarrow X_3$	+	
4	H; $X_0 \leftarrow X_1 \rightarrow X_3$	+	

Table 2. The four causal Bayesian networks being hypothesized in the example as representing the causal relationships among the three variables shown. DS(A, B) means that variables A and B are d-separated unconditionally. DS(A, B | C) means A and B are d-separated given variable C.

If alternatively, lung cancer causes smoking (row 3 of table 2) or there is a hidden cause of both lung cancer and smoking (row 4), then we could detect (from the *d*-separation patterns in of table 2) that smoking is not causing lung cancer.

Table 2 illustrates the fundamental idea underlying how constraint-based methods can discover causal knowledge from observational data, although the example is simple and limited in scope. A modest generalization of that procedure is described by Cooper (1997), where it is assumed that we have a variable like X_0 in table 2 that (among other properties) is not caused by any other measured variable; such a privileged variable has been termed an *instrumental variable* (Bowden and Turkington 1984). A major generalization of this constraint-based procedure is described, along with proofs of convergence, by Spirtes, Glymour, and Scheines (1993). Chapter 13 discusses a Bayesian method for using instrumental variables for causal inference.

Given the assumptions in section 6, is it possible to discover causal knowledge from observational data alone, without background domain knowledge? The answer is yes. If hidden variables (confounders) are excluded as possibilities, then at least three measured variables are needed. The simplest set of causal relationships that admit causal discovery from observational data is shown in figure 20. We can infer both of the following causal relationships: W causes Y, and X causes Y.

If hidden confounders may exist and we cannot assume an instrumental variable, then at least four measured variables are needed to discover a

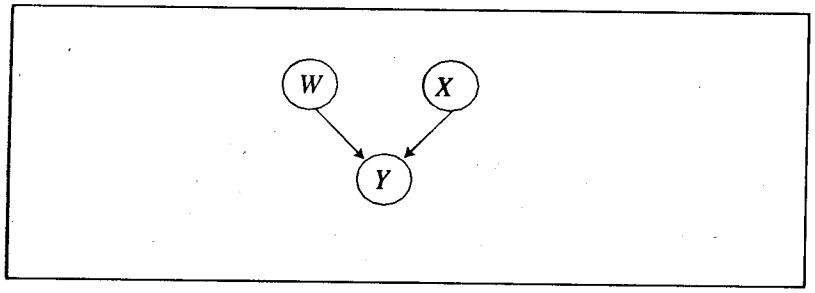

Figure 20. The simplest set of causal relationships that admit causal discovery from observational data, assuming no hidden confounders.

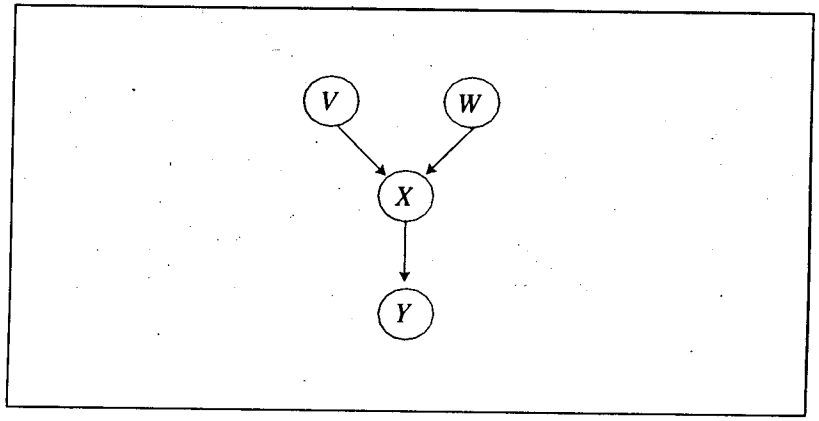

Figure 21. Simplest set of causal relationships that admit causal discovery from observational data, allowing for the possibility of hidden confounders.

causal relationship; the simplest set of causal relationships that admit causal discovery from observational data is shown in figure 21. From data generated by a causal process that can be modeled by the network shown in figure 21, we can infer only one causal relationship: X causes Y.

PC and FCI are constraint-based algorithms that considerably generalize the type of causal discovery that is illustrated in figure 21. Both algorithms are described in detail in Spirtes, Glymour, and Scheines (1993) and they have been implemented as computer programs which are commercially available (see Scheines, Spirtes, Glymour, and Meek 1995 and chapter 5). See also Pearl and Verma (1991) and chapter 3 for related research. PC and FCI assume a Bayesian network causal model, the faithfulness condition, and valid statistical testing. Current implementations of the algorithms typically involve deleting cases in which any variable has a missing value. PC assumes that hidden variables and selection bias do not exist, while FCI allows

hidden variables and selection bias. Under the assumptions that each algorithm makes, the algorithms are provably correct. PC and FCI allow the user to specify categorical background knowledge about causal relationships, as for example that one measured variable is not caused by any other measured variable. Both algorithms output a graphical pattern that expresses the causal constraints among the variables. Simple arcs express direct causality, as expected. Other edge types between nodes provide weaker constraints on the types of causal relationships between two variables, as for example an edge type indicating that the data are insufficient to resolve whether X causes Y or Y causes X. Chapter 6 describes the FCI algorithm and some of its extensions.

8. Bayesian Methods for Evaluating Causal Bayesian Networks

In 1991 Cooper and Herskovits described a general Bayesian formulation for learning causal structure (including latent variables) and parameters from observational data using Bayesian networks (Cooper and Herskovits 1992, 1991a, 1991b). To my knowledge, this was the first such description. Their Bayesian formulation assumed only that causal relationships are modeled as Bayesian networks; the basic ideas are similar to those presented later in this section. Cooper and Herskovits also specialized the general formulation by introducing a set of assumptions that make computation more tractable. Since that initial research, Bayesian causal discovery has become an active field of research in which numerous advances have been—and are continuing to be—made (Buntine 1991, 1996; Chickering and Heckerman 1996; Cooper 1995; Heckerman 1996; Heckerman, Geiger, and Chickering 1995; Meek 1995b).

Bayesian methods for causal discovery differ in several ways from constraint-based methods. First, the methods take a user-specified prior probability over Bayesian network structures and over parameters. If the user has little prior information, or it is not feasible to specify this information, then noninformative priors can be used. Given a set of modeling assumptions, the Bayesian approach combines one's prior probabilities with observational data to produce a posterior probability that conveys what one's causal beliefs should be in light of the data. Unlike constraint-based methods, no statistical testing thresholds need to be specified, but instead prior probabilities are needed.

Consider deriving the posterior probability that variable X causes variable Y given observational database D on measured variables V. Let S denote an arbitrary causal network structure containing all of the variables in V and

possibly additional hidden variables. Let K denote our background knowledge that may influence our beliefs about the causal relationships among the variables in V. Such background knowledge could come from RCEs, scientific laws, common sense, expert opinion, accumulated personal experience, as well as other sources. We can derive the posterior probability that X causes Y as

$$P(X \rightarrow Y \mid D,K) = \sum_{S:\ \{X \rightarrow Y\} \in S} P(S \mid D,K) \tag{3}$$

where the sum is taken over all causal network structures that contain an arc from X to Y and that have a nonzero prior probability. Based on the properties of probabilities, the term within the sum in equation 3 may be rewritten as follows:

$$\begin{aligned} P(S \mid D,K) &= \frac{P(S,D \mid K)}{P(D \mid K)} \\ &= \frac{P(S,D \mid K)}{\sum_S P(S,D \mid K)} \end{aligned} \tag{4}$$

Since relative to the entire set of causal structures being considered, the probability $P(D \mid K)$ is a constant, equation 4 shows that the posterior probability of causal structure S is proportional to $P(S, D \mid K)$, which we can view as a score of S in the context of D. The probability terms on the right side of equation 4 may be expanded as follows:

$$\begin{aligned} P(S,D \mid K) &= P(S \mid K)P(D \mid S,K) \\ &= P(S \mid K) \int P(D \mid S,\theta_S,K)\,P(\theta_S \mid S,K)\,d\theta_S \end{aligned} \tag{5}$$

where (1) $P(S \mid K)$ is our prior belief that S captures correctly the causal relationships among the variables in V, (2) θ_S are the probabilities (parameters) that relate the nodes in S to their parents, (3) $P(D \mid S, \theta_S, K)$ is the likelihood of data D being produced given that the causal process generating the data is isomorphic to the causal Bayesian network given by S and θ_S, and (4) $P(\theta_S \mid S, K)$ expresses our prior belief about the probability distributions that serve to model the underlying causal process. The integral in equation 5 integrates out the parameters θ_S in a Bayesian network with structure S to derive $P(D \mid S, K)$, which is called the *marginal likelihood*. Combining equations 3, 4, and 5, we obtain equation 6.

$$P(X \rightarrow Y \mid D,K) = \frac{\sum_{S:\ \{X \rightarrow Y\} \in S} P(S \mid K) \int P(D \mid S,\theta_S,K)\,P(\theta_S \mid S,K)\,d\theta_S}{\sum_S P(S \mid K) \int P(D \mid S,\theta_S,K)\,P(\theta_S \mid S,K)\,d\theta_S} \tag{6}$$

The only assumption made in equation 6 is that causal relationships are represented as Bayesian networks. Thus, the causal Markov condition is assumed. If we require that the parameter distributions given by $P(\theta_S \mid S, K)$ be continuous and contain no delta functions, then the probability of a nonfaithful distribution is infinitesimally small (Lebesgue measure zero) (see chapter 4). If we wish, however, we can express parameter priors that violate the faithfulness assumption. More likely, we might express priors that admit some distributions that are faithful and some that are not. In all cases, equation 6 will derive a valid posterior. In general, however, unfaithful distributions make identification of causal relationships more difficult for both constraint-based and Bayesian methods.

A full Bayesian approach to causal discovery considers, at least in principle, all causal models that are a priori possible. Thus, for example, the sums in equation 6 are over all possible causal structures, and the integrals are over all possible parameters for each possible causal structure. The result of such a global analysis of causality is that the derived posterior probabilities summarize a comprehensive, normative belief about the causal relationships among a set of variables.

Although equation 6 makes few assumptions, and the Bayesian theory underlying it is quite general, to render evaluation of the equation tractable, additional assumptions typically must be made, as I next describe.

One primary problem with Bayesian methods is computational tractability. Exact computation using equation 6 requires summing over a number of causal graphs that is exponential in the number of graph variables (see section 9.1). In the limited set of simulation experiments done to date, however, the application of Bayesian methods with heuristic search techniques has often been effective in rapidly and accurately recovering much of the causal generating structure on measured variables (Aliferis and Cooper 1994; Herskovits and Cooper 1990, Heckerman, Geiger, and Chickering 1995). Thus, there is hope that sometimes—perhaps even often—we can heuristically locate quickly the most probable structures that are denoted in the sums in equation 6, and then use this limited set of structures to provide a good approximation to equation 6 (Madigan and Rafferty 1994).

Under assumptions described in chapter 4, the integral in equation 6 can be computed efficiently in closed form when there are no hidden variables or missing data. When there are hidden variables or missing data, the Bayesian approach can model them explicitly and normatively; however, exact computation of the integral with current methods usually is intractable, even when causal graphs contain only a few variables. The use of sampling methods and asymptotic approximations have shown promise (Chickering and Heckerman 1996) in estimating the integral when there is missing data or hidden variables, and chapter 4 discusses several such methods.

Another challenge of applying Bayesian methods for causal discovery is the assessment of informative priors on possible causal structures and on pa-

rameters for those structures. On the one hand, the ability to represent such prior information is a great strength of the Bayesian approach. With it, we can potentially express prior causal knowledge that comes from many sources other than the observational data D. While good progress has been made in facilitating the expression of priors on Bayesian network structures and parameters (Heckerman, Geiger, and Chickering 1995), assessing such prior probabilities (particularly when there is a large set of variables) can still be difficult and sometimes infeasible; thus, assessment remains an important, open problem. Currently, it is common to specify some form of a noninformative prior on the causal structures (e.g., a uniform prior over all possible structures) and on the parameters of those structures. Noninformative priors typically require that the user specify only a few parameters; still, it sometimes is not obvious what those few parameters should be. In that case, performing a sensitivity analysis over the parameters may be a good idea.

In summary, even though exact application of Bayesian methods often is intractable, approximate solutions may be acceptable. The ability to specify structural and parameter priors is a significant strength of the Bayesian approach to causal discovery, because it allows us to incorporate into a causal analysis relevant knowledge beyond the observational data. When informative priors are not available, or are impractical to assess, noninformative priors may be used, such that the causal analysis is driven largely by the available observational data.

9. Model Search

In this section, I first describe the size of the space of causal Bayesian network structures as a function of the number of nodes in the network. Since the space is large, I provide a selected overview of methods that have been developed for searching the space.

9.1. The Size of the Model Space

Sections 6, 7, and 8 describe methods for evaluating a causal Bayesian network given a set of observational data. In this section, I describe how to use those evaluations in searching for causal models. The emphasis here is on model selection wherein we attempt to find the single best causal model that represents the relationships among the measured variables. I also briefly discuss model averaging that uses more than one model. In practice, both tasks require considering a large space of possible causal networks. In particular, as a function of the number of measured variables, the number of possible causal structures containing just those variables grows exponentially. Thus,

number of measured variables	number of causal Bayesian network structures
1	1
2	3
3	25
4	543
5	29,281
6	3,781,503
7	1.1×10^9
8	7.8×10^{11}
9	1.2×10^{15}
10	4.2×10^{18}

Table 3. The number of causal Bayesian network structures as a function of the number of variables in the network.

an exhaustive enumeration of all structures is not feasible in most domains. Table 3 shows a few sample values for the number of possible causal Bayesian network structures (i.e., directed acyclic graphs) that contain a given number of nodes; for a number of nodes greater than six, the number of structures is given in scientific notation with only two digits of accuracy in the mantissa.

The size of the space of causal Bayesian networks clearly is enormous when there are more than a few variables. Even so, it is possible that some clever algorithm could efficiently find the most probable structure, even in the worst case. However, two results suggest that in the worst case, such searches are indeed likely to be exponential time.

Bouckaert (1995) has shown that a constraint-based version of learning Bayesian networks with no hidden variables is NP-hard. Let V be a set of n binary variables and let P be a joint probability distribution over V. Assume that an oracle is available that reveals in $O(1)$ time whether a conditional independence statement holds in P. Let k be a positive constant. Consider algorithms that consult the oracle to determine whether or not there exists a causal Bayesian network structure that represents P that has at most k arcs. Bouckaert shows that this problem is NP-complete. Thus, finding the causal network structure with the minimal number of arcs is NP-hard. It can be shown that if the Markov condition and the faithfulness assumption hold, then the generating causal network contains a minimal number of arcs to represent P. Thus, if that condition and assumption hold, determining the structure of an underly-

ing causal network using constraint-based methods must be NP-hard.

Chickering (1996a) has shown that a version of Bayesian learning of Bayesian networks is NP-hard. An instance of the decision problem consists of a set of variables V, a database D, a Bayesian network structure S, the likelihood-equivalence Bayesian scoring metric $M(S, D)$ that computes equation 5 (see chapter 4 for a definition of likelihood equivalence), and a real value p. The decision question is as follows: Does there exist a network structure S defined over the variables in V, where each node in S has at most k parents (for k greater than 1), such that $M(S, D) \geq p$? Chickering shows that this problem is NP-complete. Thus, finding the causal network structure with the maximum score is NP-hard. The proof of NP-completeness uses a reduction that relies on informative priors.

Thus far in this section I have considered Bayesian networks on measured variables only. Chapter 4 discusses several methods for estimating equation 5 when S contains hidden (latent) variables or missing data.

9.2. Search Algorithms

Since searching the usually enormous space of causal Bayesian networks appears infeasible, researchers have developed a number of approaches to cope with the task. In the remainder of this section I provide a brief survey of a selected set of those approaches. No attempt is made to provide complete coverage of all the algorithms that have been developed. As concrete examples, three search algorithms (PC, K2, and OccamsWindow) are described in more detail than the others.

9.2.1. Search Algorithms for Constraint-Based Causal Discovery

In this section, I provide a brief summary of the PC search algorithm that was developed by Spirtes, Glymour, and Scheines (1993) (figure 22). In figure 22, steps 1 and 4 are performed in $O(n^2)$ time. PC has relatively efficient techniques for performing steps 2 and 3. In the worst case, however, steps 2 and 3 require time that is exponential in n. The worst cases occur when the nodes in the generating graph are highly connected, and therefore, F is a dense graph. In particular, the computational time complexity of PC is bounded from above by

$$n^2(n-1)^{k-1}/(k-1)!$$

where k is the maximum number of edges directly connected to a node in graph F that is produced by Step 2 (Spirtes, Glymour, and Scheines 1993). Thus, when k is bounded, the complexity is polynomial in the number of nodes. This analysis provides a loose upper bound on the worst-case time complexity; the expected time complexity will depend on the details of the underlying causal model and the data it produces.

procedure PC;

{Input: A set of n nodes, and a function T to test conditional independence of sets of nodes.}

{Output: A set of arcs that indicate causal relationships between variables, and a set of undirected edges that indicate relationships between variables in which causal directionality is left undetermined.}

{Assumptions: The data generating process is a causal Bayesian network, the faithfulness assumption holds, the test T is correct, and there are no missing data or hidden variables.}

Step 1. Form a complete undirected graph C on the n nodes.

Step 2. Using T, begin with low order conditional independence tests and progressively remove edges from C whenever two nodes are marginally or conditionally independent. Let F denote the resulting undirected graph.

Step 3. For each triple of nodes (X, Y, Z) such that (X, Y) and (Y, Z) are each adjacent in F, but (X, Z) is not, do the following: Orient $X - Y - Z$ as $X \rightarrow Y \leftarrow Z$ if and only if X and Z are dependent when conditioned on each subset of the nodes (excluding X and Z) that contains Y. Let F' denote the resulting partially directed graph.

Step 4. Repeat, until no more edges in F' can be oriented:

a. If in F' it is the case that $X \rightarrow Y$ appears, $Y - Z$ appears, and X and Z are not connected (by an undirected edge or an arc), then orient $Y - Z$ as $Y \rightarrow Z$.

b. If there is a directed path from X to Y, and $X - Y$ is in F', then orient $X - Y$ as $X \rightarrow Y$.

end {PC};

Figure 22. The PC algorithm.

The output of Step 4 in PC is called a *pattern*, which is a term introduced by Verma and Pearl (1990). Arcs represent causal relationships. Undirected edges indicate relationships in which causal directionality is left undetermined by PC. The PC algorithm does not always orient all undirected edges that can be oriented. For a discussion of rules for obtaining a complete orientation, see Meek (1995), as well as the discussion of essential graphs in section 9.2.2.

Figure 23 shows an example of applying PC. In this example, I assume that the causal network that generated the data is shown at the top of figure 23. In Step 1 in the figure all four nodes are connected by edges. In Step 2, A and B are marginally independent, so the edge between them is removed. Also in Step 2, A and D are independent given C, and therefore the edge between them is removed. Similarly, B and D are independent given C. The final graph in Step 2 contains just three edges, which form a skeleton indicating variables that have direct causal relationships. In Step 3 of the example, it happens that the independence conditions on A, B, and C that fol-

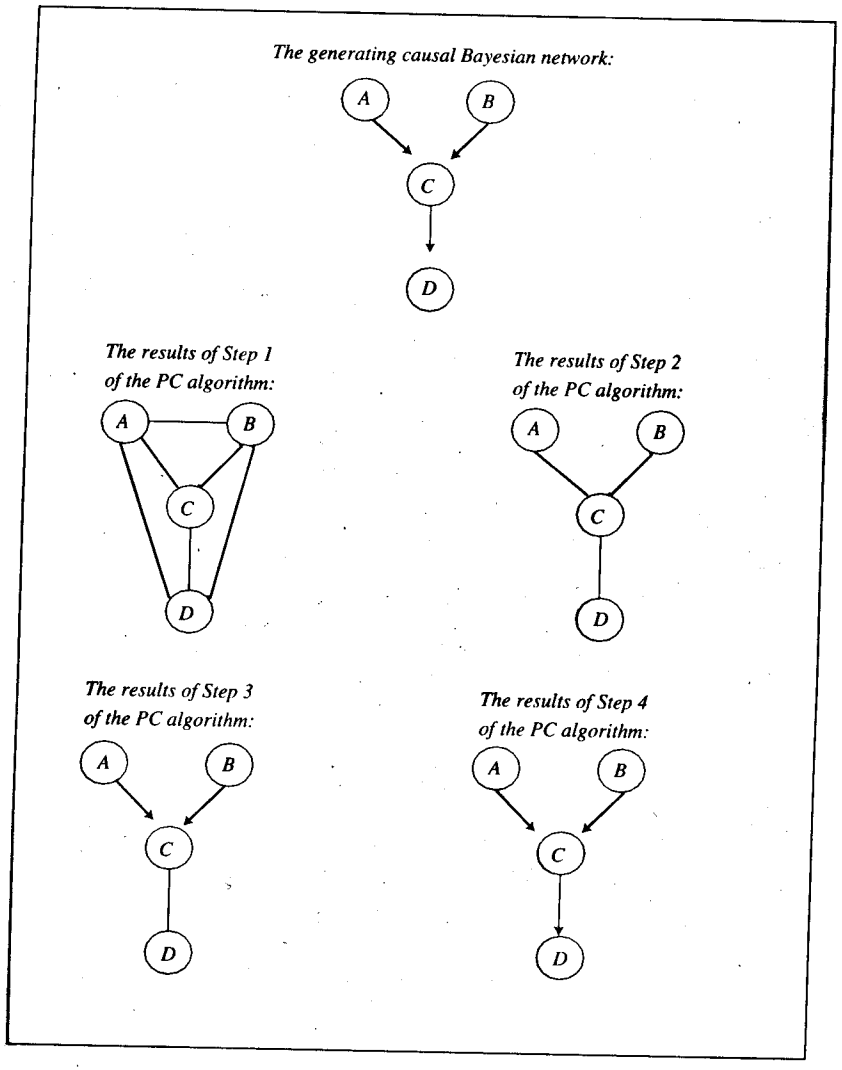

Figure 23. An example of the application of the PC algorithm.

low from the generating network are consistent only with an arc from A to C and an arc from B to C. No other arcs can be generated in Step 3. In Step 4a we orient an arc from C to D and quit.

In figure 23, I have shown one of the simplest possible applications of PC. For more complex applications of PC, see Spirtes, Glymour, and Scheines (1993), which also describes extensions to the algorithm. As mentioned in

section 7, FCI is a constraint-based causal discovery algorithm that admits both measured and hidden variables. A detailed description and analysis of FCI also is given by Spirtes, Glymour, and Scheines (1993).

9.2.2. Search Algorithms for Bayesian Causal Discovery

In this section I describe several heuristic search algorithms that have been used along with Bayesian scoring metrics to search for the most probable causal network given a set of observational data. Many of these algorithms have close parallels to search algorithms used in statistics for constructing predictive models (e.g., logistic regression models), although I do not focus here on that comparison. I also describe a method for model averaging. All the algorithms I describe use $P(S, D \mid K)$ from equation 5 as a *scoring metric* to rank causal network structures, since $P(S, D \mid K)$ is proportional to $P(S \mid D, K)$ given a database D.

Special Case Algorithms. Researchers have developed special case search algorithms that are efficient for restricted causal network structures. When we can assume that each node in the generating network has at most one parent, then a polynomial time algorithm exists for finding the most probable structure (or set of structures) (Heckerman, Geiger, and Chickering 1995). Unfortunately, such restrictions rarely apply, and thus the need for other search methods.

Greedy Search Algorithms. Greedy search algorithms work by adding, removing, and/or reversing a few arcs (typically one) at each step of the search. The search halts when there is no greedy step that improves the scoring metric. A forward stepping algorithm, for example, typically begins with a network that contains no arcs, and then it adds incrementally that arc whose addition most increases the probability of the resulting structure. When the addition of no single arc can increase the scoring metric, the algorithm stops adding arcs. A backward stepping algorithm usually begins with a fully connected network and then removes one arc at a time, until no single arc can be removed to increase the scoring metric. Combinations of forward and backward stepping algorithms have been developed as well (Heckerman, Geiger, and Chickering 1995; Spirtes and Meek 1995).

Some greedy search algorithms have assumed a causal ordering of the nodes, such that the potential parents of a node X are just the nodes that are lower in the ordering than X (Cooper 1990). Time precedence, for instance, sometimes can provide such information; in general, however, an ordering will not be available. Researchers have therefore explored greedy algorithms that do not require an ordering (Buntine 1991; Heckerman, Geiger, and Chickering 1995). A problem with greedy search algorithms is they may halt at local maxima of the scoring metric, rather than at a global maxima (Xiang, Wong, and Cercone 1996) Various procedures, such as multiple search restarts from random graphs, can be used in an attempt to ameliorate the lo-

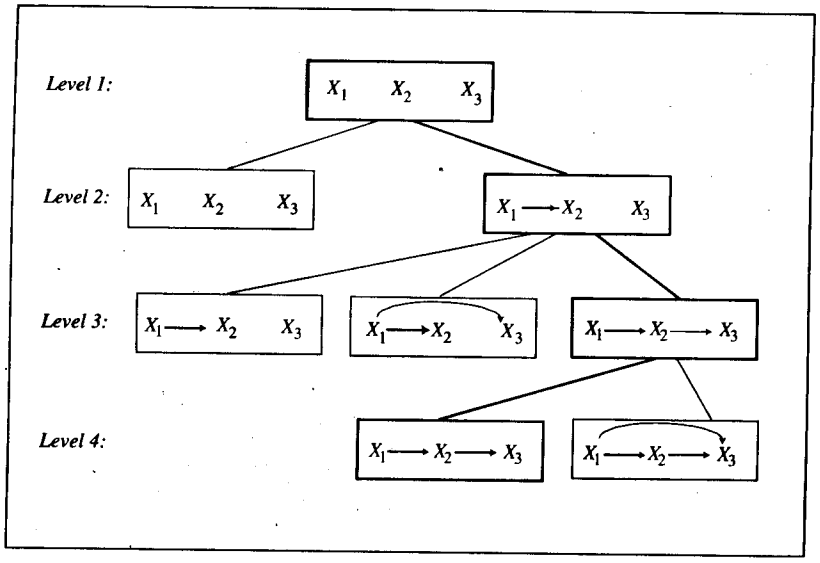

Figure 24. An example of the application of the K2 forward stepping search algorithm. The path taken in the search is shown in bold.

cal maxima problem by taking the maximum of a set of local maxima.

I now illustrate a simple application of a forward stepping search algorithm called K2 (Cooper and Herskovits 1992). The K2 algorithm makes assumptions that allow it to search for the parents of each node separately. Figure 24 shows an example in which the search starts at level 1 with no arcs. Suppose the node ordering is given by the list (X_1, X_2, X_3), so that the potential parents of a node are given by the nodes to its left in the list. The search starts with the first node in the list. At level 1, K2 searches for the parents of node X_1, which can have no parents, according to the node ordering. So, the search advances to level 2. Here the algorithm searches for the parents of X_2 by considering no arcs (left box) and one arc (right) from X_1. Suppose the causal network structure on the right has the highest score, where the score for a structure S is given by $P(S, D \mid K)$. Since there are no additional arcs to consider as possible parents of X_2, the algorithm fixes X_1 as a parent of X_2 (shown as a bold arc) and then it continues onward to consider X_3. For node X_3, K2 considers each possible single arc addition, of which there are two, as well as the addition of no arc. Suppose the rightmost causal network structure receives the highest score. In that case, K2 fixes X_2 as a parent of X_3. Next, it considers all additional single arcs it could add into X_3, of which there is only one, which is from X_1. Suppose

the rightmost structure at level 4 has a score lower than $X_1 \rightarrow X_2 \rightarrow X_3$. In that case the algorithm halts and returns $X_1 \rightarrow X_2 \rightarrow X_3$.

Greedy Search Algorithms on Essential Graphs. Markov equivalence is a relationship based on independence that establishes a set of equivalence classes of causal network structures relative to a set of measured variables. In particular, these structures are statistically indistinguishable based on independence relationships among the measured variables. Let U be a set of causal network structures representing one such equivalence class. If for every S in U, an arc between two nodes is oriented only in one direction, then retain the arc; otherwise replace it with an undirected edge. The resulting graph was independently developed and investigated by several researchers (Andersson, Madigan, and Perlman 1997; Chickering 1995; Meek 1995), who use the terms *maximally oriented graph, completed pdag representation,* and *essential graph* to describe it. This chapter uses the term essential graph. Essential graphs are a special case of chain graphs (Wermuth and Lauritzen 1990), both of which can contain arcs as well as undirected edges. For example, consider a two node model with variables X and Y. There are three causal network structures that contain just these two measured variables (X no_arc Y, $X \rightarrow Y$, and $X \leftarrow Y$), but only two essential graphs (X no_arc Y and $X - Y$) because in this case $X \rightarrow Y$ and $X \leftarrow Y$ are Markov equivalent.

Researchers have developed greedy algorithms for searching over essential graphs (Anderson, Madigan, and Perlman 1998; Chickering 1996; Meek 1995; Spirtes and Meek 1995). For a given set of n measured variables, there are fewer essential graphs than causal network structures (directed acyclic graphs). For example, for n equal to 4, there are 543 causal network structures, but only 185 essential graphs.[4] Since the essential graph space is smaller, it is potentially easier to search. Thus, searching the space of essential graphs appears to be a promising method for causal Bayesian network selection.

Model Averaging. Suppose we perform an inference, as for example to derive the probability that in light of observation U the manipulation of variables W and X to particular values will cause variable Y to have a particular value. The normative Bayesian approach to performing this inference is model averaging. Model averaging involves performing the inference for each possible causal model and multiplicatively weighting the inference result (a probability) by the posterior probability of the causal model. The Bayesian inference is just the sum of these weighted inferences.

The large space of possible causal network structures generally makes it infeasible to perform complete model averaging. Researchers, therefore, have investigated heuristic methods for performing selective model averaging. These methods heuristically search for high probability causal network structures; the networks encountered during the search are used for model averaging.

Madigan and colleagues developed an algorithm called Occam's Window

procedure OccamsWindow;

Step 1. Initialize a set A of models {e.g., place in A a Bayesian network with no arcs}

Step 2. Consider each possible legal, one-step, greedy modification to each member of A. Modifications include single arc additions, deletions, and reversals that induce no cycles.

Step 3. Choose the modification in Step 2 that leads to the highest scoring structure S (if any) that satisfies the following conditions:

- The highest scoring structure already in A is not more than 20 times greater than $P(S, D \mid K)$.
- There is no model S' in A that is a subgraph of S such that $P(S', D \mid K) > P(S, D \mid K)$.

Step 4. If step 3 produces a new structure S, then add S to A and go to Step 2; otherwise, continue to step 5.

Step 5. Perform model averaging using the structures in A.

end; {OccamsWindow}

Figure 25. The Occam's Window algorithm.

(Madigan and Raftery 1994), which can be applied to search for a set of causal Bayesian network structures with which to do model averaging. A high-level description of the algorithm is shown in figure 25. Researchers also have investigated a method for performing heuristic model averaging using essential graphs rather than Bayesian networks (Madigan, Andersson, Perlman, and Volinsky 1996).

Combining Constraint-based and Bayesian Methods. Researchers have developed hybrid search algorithms that have two stages (Singh and Valtorta 1993, Spirtes and Meek 1995). The first stage involves selecting a causal network structure using a constraint-based search method. The second stage involves using that structure to start a Bayesian search. The main idea underlying these algorithms is that constraint-based methods provide a relatively quick first approximation structure, which is then refined by the Bayesian methods.

Other combinations of constraint-based and Bayesian methods are possible, and the general approach appears promising. For instance, chapter 14 describes a method that uses a constraint-based method for constructing the structure of a causal model and a Bayesian method for parameter estimation. As another example, as mentioned in section 9.1, an unconstrained search for hidden variables using a Bayesian scoring method generally is intractable. It may be more effective to use constraint-based methods to limit the space of

latent variable networks that are searched. Recently researchers have investigated a latent-variable version of the essential graph search method mentioned earlier in this section (Spirtes, Richardson, and Meek 1997). In this recent research, a representation called a partial ancestral graph (PAG) is used to model equivalence classes of causal Bayesian networks that may include latent variables. A greedy search is performed in the space of PAGs. For each greedy step, the corresponding PAG is converted to a mixed ancestral graph (MAG), which is a completely oriented PAG. The MAG is then evaluated using a Bayesian information criterion (BIC) scoring metric, which can be viewed as an asymptotic approximation to a Bayesian scoring metric. The MAG with the highest score is converted back to a PAG, and the search continues until no local addition to the leading PAG improves the BIC score. Starting with that leading PAG, a similar procedure is applied using backward stepping search. When backward searching is completed, the final, leading PAG is returned. Preliminary simulation experiments, while very limited in scope, are encouraging (Spirtes, Richardson, and Meek 1997). The use of a Bayesian scoring metric to search an abstraction space of causal network models (e.g., PAGs) appears to be a promising direction of research.

10. A Selected Summary of Prior Results

Inducing causal relationships from observational data is challenging for many reasons, including the possible presence of latent variables, selection bias, missing data, limited amounts of records, a large search space, and statistical indistinguishability. Even so, as this chapter has shown, there are circumstances in which we can induce causal relationships from observational data. Characterizing all the circumstances and their consequences regarding causal discovery remains an open problem. Nonetheless, presently we do have some useful understanding of the conditions that make causal discovery possible. While it is beyond the scope of this chapter to provide a complete survey of present understanding, in the remainder of section 10 I provide a selected summary of some key results.

10.1. Convergence Results

The PC algorithm assumes the Markov condition, the faithfulness condition, valid statistical testing, no latent variables, and no selection bias. Under these assumptions, in the large sample limit, PC will recover all causal relationships that can be recovered from observational data. Bouckaert (1995) has shown that under those assumptions Bayesian learning methods will recover the same causal relationships as PC in the large sample limit. This result as-

sumes that the Bayesian method does not include prior probabilities of 0 or 1 on structures or parameters, and that model search is exhaustive.

The FCI constraint-based discovery method models latent variables and selection bias. In the large sample limit, FCI has been proved to converge to a model that contains no incorrect causal statements if the modeling assumptions it makes are correct (Spirtes, Glymour, and Scheines 1993). Within the FCI language for representing causal constraints, it also has been shown that the algorithm is complete (Spirtes, Glymour, and Scheines 1993). The degree to which the FCI language itself is complete, relative to all possible causal constraints that can be expressed, remains an open question. There are known examples in which independence constraints alone are insufficient to discover a causal constraint that can be discovered from observational data (Verma and Pearl 1990). As a simple example, since FCI does not model the number of values of latent variables, it clearly is unable to learn what may be learnable about that number. Bayesian methods, conversely, are able to model and learn about the number of values of hidden variables (Chickering and Heckerman 1996; Cooper 1995; Cooper and Herskovits 1992; Geiger, Heckerman, and Meek 1996). Bayesian methods, however, are computationally demanding, although as mentioned in section 8, approximation methods appear promising (see chapter 4). Current Bayesian methods evaluate a specific causal Bayesian network, or somewhat more generally, an essential graph that represents an equivalence class of Bayesian networks. More abstract causal constraints (e.g., X either causes Y or causes Z) can be constructed from the evaluation of a set of causal Bayesian networks. It remains an open problem, however, to investigate the extent to which abstract constraints can be evaluated (scored) directly using Bayesian methods.

10.2. Simulation Studies

In this section, I describe the results of studies in which a database of cases was generated from a Bayesian network by simulation and then given as input to an algorithm that attempted to discover causal relationships. Since the ALARM Bayesian network has been widely used for simulation studies, I focus on experiments that have used data generated from that network.

Beinlich constructed the ALARM network as a research prototype to model potential anesthesia problems in the operating room (Beinlich, Suermondt, Chavez, and Cooper 1989; Cooper and Herskovits 1992). ALARM contains 46 arcs and thirty-seven nodes, and each node has from two to four possible values.

Cases were generated from ALARM using a Monte Carlo simulation technique (Henrion 1988). Although all the studies mentioned here have used the same network structure for ALARM, variations of the probability parameters have been applied. Each ALARM case corresponds to a value assignment to

each of the thirty-seven variables. The simulation technique is an unbiased generator of cases, in the sense that the probability that a particular case is generated is equal to the probability of the case according to the Bayesian network.

Cooper and Herskovits applied K2 with a database of 3,000 ALARM cases (Cooper and Herskovits 1992). K2 also was given an ordering on the thirty-seven nodes that is consistent with the partial order of the nodes as specified by the ALARM network. From the 3,000 cases, K2 constructed a network identical to ALARM, except that one arc was missing and one arc was added. A subsequent analysis revealed that the missing arc is not strongly supported by the 3,000 cases. The extra arc was added because of the greedy nature of the search algorithm. Total search time was approximately five minutes when using a circa 1990 personal computer.

Heckerman, Geiger, and Chickering (1995) developed and investigated a greedy search algorithm that generalized K2 by removing the assumption of a node ordering. A Bayesian scoring metric similar to that used by K2 was applied as well. Although the algorithm was able to estimate the ALARM joint distribution accurately, as measured by cross entropy, the estimated structural model had on average forty-five arc differences from the ALARM network; arc difference is defined as

$$\sum_{i=1}^{n} \delta_i ,$$

where δ_i is the symmetric difference of the parents of node x_i in ALARM and the parents of x_i in the learned network. This difference is conservative because it counts arcs that are reversed from ALARM, even when those arc orientations are statistically indistinguishable. When a much slower simulated annealing search algorithm was applied to the same data set, only about twenty differences existed on average.

Algorithms that perform a greedy search over a space of essential graphs have been applied to ALARM data sets (Chickering 1996; Meek 1998; Spirtes and Meek 1995). The best results to date of any search algorithm on ALARM data has been achieved by Meek's greedy equivalence search (GES) algorithm (Meek 1998). The GES algorithm first performs a forward stepping greedy search followed by a backward stepping greedy search. When given a data set containing 10,000 ALARM cases, and run on a UNIX workstation, the algorithm returns in about 5 hours a network that contains only one error, namely, a missing edge between two variables that are only very weakly statistically associated in ALARM. When an arc and its reversal are statistically indistinguishable in the generating ALARM network, and one of those arc orientations is in ALARM, the evaluation of GES counted the presence of either orientation as acceptable (no error). These results are impressive and suggest that the GES algorithm deserves considerably more study.

Chapter 6 describes a preliminary experiment in which 10,000 cases gen-

erated from ALARM were used to evaluate the FCI algorithm when there are latent variables. Selection bias also was examined, but those results are not summarized here. The primary metrics of evaluation were the percentage of ancestor and nonancestor relationships correctly predicted, according to the ALARM network. A node X is an ancestor of node Y if there is a directed path from X to Y. When there were no latent variables, FCI correctly predicted all of the ancestor relationships (pairs) and 97 percent of the nonancestor relationships. When node 29 was considered a latent variable, FCI correctly predicted all ancestor relationships and 91 percent of the nonancestor relationships. When both nodes 29 and 22 were considered as latent variables, FCI still correctly predicted all the ancestor relationships and 92 percent of the nonancestor relationships. Although these results provide useful insight into the performance of FCI, we have much more to learn about how causal discovery algorithms perform when there are latent variables.

In this section, I have given only a sampling of simulation results that have been reported. Other simulation studies include an extensive set of experiments using constraint-based methods that are described in Spirtes, Glymour, and Scheines (1993), and experiments applying K2 to random graphs, as described by Aliferis and Cooper (1994). I conclude this section with a summary of the latter results.

Aliferis and Cooper generated sixty-seven Bayesian networks in a randomized fashion (see Aliferis and Cooper 1994 for details), such that each network contained from two to fifty nodes, two or three values per node, and zero to ten parents per node. For each network, the number of cases generated was randomly (uniformly) selected to be in the range from 0 to 2,000. The probability parameters for each network also were randomly generated. The K2 search algorithm and metric were applied to each of the sixty-seven datasets, where K2 was given a node ordering consistent with the generating Bayesian network. In brief, on average K2 found 92 percent of all the arcs in the generating network and erroneously added 5 percent more arcs than existed in the generating network.

In experiments based on simulated databases, we know the causal network that generated the data, and thus, we know the underlying causal reality. We therefore can judge the causal discovery performance of an algorithm relative to that generating causal network. There are, however, two major weaknesses of such studies. First, we assume the existence of a causal process that can be modeled by a Bayesian network. Thus, we evaluate the internal validity of discovery methods relative to an assumed model of causality. We are not testing external validity relative to the real world. Second, assuming a Bayesian network model, we still need to parameterize the network with a set of probabilities that are used to stochastically generate a database. It can be difficult to know what these probabilities should be, particularly in the presence of latent variables. One strategy is to choose random probabilities

to parameterize a given Bayesian network structure. As a form of sensitivity analysis, we can generate multiple random parameterizations, and for each one we generate a database that is used to evaluate a causal discovery algorithm. Unfortunately, it may be highly unlikely that a given random parameterization will closely resemble any real-world causal process.

An arguably better approach is to have an expert generate a causal Bayesian network based on personal knowledge and knowledge from the literature. The ALARM network is one example. Such networks have the advantage that we know their structure precisely and their probabilities are likely to resemble those of the real-world processes being modeled. One disadvantage of the approach is that manually generating such networks is labor intensive. Another disadvantage is that it relies on human knowledge of a causal process, which may be incomplete or incorrect.

A related, and relatively unexplored approach, would be to use causal Bayesian networks that represent human-constructed entities (e.g., a jet engine). For example, Spirtes, Glymour, and Scheines (1993, p. 243) describe such a study in which the PC algorithm correctly identifies the subcomponents of a qualification test taken in the military. Such experiments do not, of course, inform us directly about the performance of causal discovery methods when using data on natural systems.

10.3. Studies Using Real Databases

Using observational data, we would like to discover new and useful causal knowledge about the real world. Evaluating the performance of causal discovery methods in achieving that goal is significantly more difficult than evaluating most statistical and machine-learning algorithms, such as classification algorithms. The difficulty stems from being unable to simply use a test set of reserved cases for evaluation. Since manipulation is intrinsic to the notion of causality (at least as I use the term *causality*), the evaluation of a causal hypothesis involves knowledge of what follows from such a manipulation. Sources of such knowledge include expert subjective knowledge and RCEs, which I now discuss in turn.

Several studies have examined causal discovery that is evaluated based on human judgment, which often is rendered in an original paper in the literature. Examples include causal discovery in the areas of publication productivity, education and fertility, American occupational structure, the influences on college plans, and abortion opinions (Spirtes, Glymour, and Scheines 1993). Although an account of these studies is beyond the scope of this chapter, on the whole the causal relationships discovered often (but not always) are consistent with human judgment. Where there is deviation, we generally do not know whether it is due to incorrect algorithmic output, inadequate human knowledge, or both.

Ideally we would have available the results of large, well-performed RCEs to validate causal relationships that are suggested from observational data by causal discovery algorithms. Conducting RCEs can be problematic for the reasons outlined in section 2. More feasibly, we might look for prior studies in which observational data was obtained on a set of variables in a given domain and RCE data was obtained on the same or a similar set of variables in the domain. Preferably, the context and entity population sample would be exactly the same, but an informative study could certainly occur in the absence of such an ideal. If the causal relationships suggested using an observational database coincide closely with the causal relationships suggested by the corresponding RCE, then we have positive support for methods for causal discovery from observational data. If such comparisons were carried out over many databases and domains, we would begin to get a clear picture of the strengths and weaknesses of present methods for causal discovery from observational data.

Spirtes, Glymour, and Scheines (1993) use the discovery methods described in this book to analyze observational biological data of Spartina biomass. They applied the PC algorithm to field data collected by Linthurst (1979). Linthurst collected observational data on 14 possible factors that influence the growth of Spartina biomass, as well as the actual biomass. At each of nine sites in the Cape Fear Estuary five data samples were measured. Using this data, the PC algorithm output only pH as a cause of biomass. Original laboratory experiments performed by Linthurst showed pH to be a causal factor influencing biomass. The experiments also showed salinity to be a causal factor at fairly neutral pH levels that were sparsely represented in the field data. Aeration did not significantly influence biomass in the laboratory. Linthurst did not experimentally examine all the other 11 possible factors, so we do not know whether any of them would show a causal effect on biomass. Overall, the known experimental results support the output of the PC algorithm as being essentially correct for the Spartina biomass domain.

As another example, chapter 15 describes a study that involves causal models found by a constraint-based discovery algorithm that are compared to current knowledge and theory of gas exchange in plant leaves. The results reveal some strengths and weaknesses of the discovery method for this domain.

Clearly, more studies are needed that compare the causal relationships derived from observational investigations with those derived from experimental investigations. As researchers locate parallel observational and experimental databases, additional studies will be possible.

Acknowledgments

I thank Constantin Aliferis, John Aronis, Clark Glymour, David Heckerman, Chris Meek, Peter Spirtes, and Stefano Monti for helpful comments on earlier drafts of this chapter. The writing of this chapter and portions of the research presented here were supported in part by grants BES-9315428 and IRI-9509792 from the National Science Foundation and by grant LM05291 from the National Library of Medicine. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or the National Library of Medicine.

Notes

1. A *graph* consists of *nodes* (typically represented as circles—in the structural equation modeling literature [Bollen 1989], usually measured variables are represented with squares and unmeasured variables with circles; in the causal network structures in the current chapter, we simply use circles for all variables) with *edges* between some pairs of the nodes. Nodes are also called vertices. If every edge has a direction associated with it (typically represented with an *arc* $\rightarrow$), the graph is a *directed graph*. A *directed path* from node X to node Y is a sequence of nodes, beginning with X and ending with Y, such that there is an arc from each node to its successor in the sequence. A directed path from a node to itself is called a *directed cycle*. A directed graph that contains no directed cycles is called a *directed acyclic graph*. A *parent* of node X is a node W for which there is an arc from W to X. Node X is then said to be a *child* of W. If there is a directed path from V to X, then V is an *ancestor* of X. Correspondingly, X is said to be a *descendant* of V. The *nondescendants* of a node X are all the other nodes in the graph that are not descendants of X. A subgraph G' of a graph G is a graph that contains just a subset of the nodes in G and all the edges in G among that subset of nodes.

2. The notation $manipulate_i(\mathbf{X}')$ corresponds to the same manipulations as $manipulate_i(\mathbf{X})$ for that subset of $\mathbf{X}$ represented by $\mathbf{X}'$. Similarly, $manipulate_i(\mathbf{X}'')$ corresponds to the same manipulations as $manipulate_i(\mathbf{X})$ for that subset of $\mathbf{X}$ represented by $\mathbf{X}''$. An analogous correspondence holds for $manipulate_j(\mathbf{X})$, $manipulate_j(\mathbf{X}')$, and $manipulate_j(\mathbf{X}'')$.

3. With a smooth distribution, we exclude discontinuities and delta functions in the probability density functions.

4. Joel Martin provided this count of essential graphs, based on exhaustive enumeration (personal communication, March 1994).

References

Aliferis, C. F. 1998. A Temporal Representation and Reasoning Model for Medical Decision-Support Systems, Ph.D dissertation, University of Pittsburgh, Intelligent Systems Program, Pittsburgh, Penn.

Aliferis, C. F., and Cooper, G. F. 1998. Causal Modeling with Modifiable Temporal Belief Networks. Center for Biomedical Informatics. Technical Report 01, University of Pittsburgh, Pittsburgh, Penn.

Aliferis, C. F., and Cooper, G. F. 1995. A New Formalism for Temporal Modeling in Medical Decision-Support Systems. In *Proceedings of the Symposium on Computer Applications in Medical Care*. Philadelphia, Penn: Hanley and Belfus.

Aliferis, C. F., and Cooper, G. F. 1994. An Evaluation of an Algorithm for Inductive Learning of Bayesian Belief Networks Using Simulated Data Sets. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 8–14. San Francisco: Morgan Kaufmann Publishers.

Andersson, S. A.; Madigan, D.; and Perlman, M. D. 1997. A Characterization of Markov Equivalence Classes for Acyclic Digraphs. *Annals of Statistics.* 25(2): 505-541.

Angrist, J. D.; Imbens, G. W.; and Rubin, D. B. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91(434): 444– 472.

Balke, A., and Pearl, J. 1994. Counterfactual Probabilities: Computational Methods, Bounds, and Applications. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 46–54. San Francisco: Morgan Kaufmann Publishers.

Beinlich, I. A.; Suermondt, H. J.; Chavez, R. M.; and Cooper, G. F. 1989. The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks. Paper presented at the Second European Conference on Artificial Intelligence in Medicine, London, August.

Berzuini, C.; Bellazi, R.; Quaglini, S.; and Spieglehalter, D. J. 1992. Bayesian Networks for Patient Monitoring. *Artificial Intelligence in Medicine* 4(3): 243–260.

Bollen, K. A. 1989. *Structural Equation Models with Latent Variables.* New York: Wiley.

Bouckaert, R. 1995. Bayesian Belief Networks: From Construction to Inference, Ph.D. dissertation, Computer Science Department, University of Utrecht, Netherlands.

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

Bulpitt, C. J. 1996. *Randomized Controlled Clinical Trials.* Norwell, Mass.: Kluwer Academic.

Buntine, W. 1996. A Guide to the Literature on Learning Probabilistic Networks from Data. *IEEE Transactions on Knowledge and Data Engineering* 8(2): 1–17.

Buntine, W. 1991. Theory Refinement in Bayesian Networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 52–60. San Francisco: Morgan Kaufmann Publishers.

Cartwright, N. 1989. *Nature's Capacities and Their Measurement.* New York: Oxford University Press.

Castillo, E.; Gutierrez, J. M.; and Hadi, A. S. 1997. *Expert Systems and Probabilistic Network Models.* New York: Springer-Verlag.

Chickering, M. 1996a. Learning Bayesian Networks Is NP-Complete. In *Learning from Data: Lecture Notes in Statistics,* eds. D. Fisher D. and H. Lenz, 121–130. New York: Springer- Verlag.

Chickering, D. M. 1996b. Learning Equivalence Classes of Bayesian Network Structures. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 150–157. San Francisco: Morgan Kaufmann Publishers.

Chickering, M. 1995. A Transformational Characterization of Equivalent Bayesian Network Structures. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 87–98. San Francisco: Morgan Kaufmann Publishers.

Chickering, D. M., and Heckerman, D. 1996. Efficient Approximations for the Marginal Likelihood of Incomplete Data Given a Bayesian Network. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 158–168. San Francisco: Morgan Kaufmann Publishers.

Cooper, G. F. 1997. A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships. *Journal of Data Mining and Knowledge Discovery* 1(2): 203–224.

Cooper, G. F. 1995a. A Method for Learning Belief Networks That Contain Hidden Variables. *Journal of Intelligent Information Systems* 4 (4):1–18.

Cooper, G. F. 1995b. Causal Discovery from Data in the Presence of Selection Bias. Paper presented at the International Workshop on Artificial Intelligence and Statistics, January 4–7, Ft. Lauderdale, Florida.

Cooper, G. F. 1990. The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks. *Artificial Intelligence* 42(2–3): 393–405.

Cooper, G. F., and Herskovits, E. 1992. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* 9(4): 309–347.

Cooper, G. F., and Herskovits, E. H. 1991a. A Bayesian Method for Constructing Bayesian Belief Networks from Databases. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 86–94. San Francisco: Morgan Kaufmann Publishers.

Cooper, G. F., and Herskovits, E. H. 1991b. A Bayesian Method for the Induction of Probabilistic Networks from Data, SMI-91-1, Section of Medical Informatics, University of Pittsburgh, Pittsburgh, Penn.

Dagum, P., and Galper, A. 1993. Forecasting Sleep Apnea with Dynamic Network Models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 64–71. San Francisco: Morgan Kaufmann Publishers.

Dagum, P., and Luby, M. 1993. Approximating Probabilistic Inference in Bayesian Belief Networks Is NP-Hard. *Artificial Intelligence* 60(1): 141–153.

Friedman, L. M.; Furberg, C. D.; and DeMets, D. L. 1996. *Fundamentals of Clinical Trials.* 3d ed. St. Louis, Mo.: Mosby.

Geiger, D.; Heckerman, D.; and Meek, C. 1996. Asymptotic Model Selection for Directed Networks with Hidden Variables. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 283–290. San Francisco: Morgan Kaufmann Publishers.

Geiger, D.; Verma, T.; and Pearl, J. 1990. Identifying Independence in Bayesian Networks. *Networks* 20(5): 507–534.

Heckerman, D. 1996. A Tutorial on Learning with Bayesian Networks, MSR-TR-95-06, Microsoft Research, Redmond, Wash. Available at http://www.research.microsoft.com/research/dtg/heckerma/heckerma.html

Heckerman, D., and Shachter, R. 1995. A Definition and Graphical Representation of Causality. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 262– 273. San Francisco: Morgan Kaufmann Publishers.

Heckerman, D.; Geiger, D.; and Chickering, D. 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20(3): 197–243.

Heise, D. R. 1975. *Causal Analysis.* New York: Wiley.

Henrion, M. 1990. An Introduction to Algorithms for Inference in Belief Nets. In *Uncertainty in Artificial Intelligence 5*, eds. M. Henrion, R. D. Shachter, L. N. Kanal, and J. F. Lemmer, 129–138. Amsterdam, The Netherlands: North-Holland.

Henrion, M. 1988. Propagating Uncertainty in Bayesian Networks by Logic Sampling. In *Uncertainty in Artificial Intelligence 2*, eds. J. F. Lemmer and L. N. Kanal, 149–163. Amsterdam, The Netherlands: North-Holland.

Herbert, N. 1985. *Quantum Reality*. Garden City, N.Y.: Anchor.

Herskovits, E. H., and Cooper, G. F. 1991. KUTATO: An Entropy-Driven System for the Construction of Probabilistic Expert Systems from Databases. In *Uncertainty in Artificial Intelligence 6*, ed. P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer. Amsterdam: Elsevier North Holland.

Howard, R. A., and Matheson, J. E. 1984. Readings on the Principles and Applications of Decision Analysis, Strategic Decisions Group, Menlo Park, California.

Jensen, F. V. 1996. *An Introduction to Bayesian Networks*. New York: Springer-Verlag.

Kim, J. H., and Pearl, J. 1983. A Computational Model for Combined Causal and Diagnostic Reasoning in Inference Systems. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 190–193. San Francisco: Morgan Kaufman.

Lam, W., and Bacchus, F. 1994. Learning Bayesian Belief Networks, An Approach Based on the MDL Principle. *Computational Intelligence* 10(3): 269–293.

Linthurst, R. A. 1979. Aeration, Nitrogen, pH, and Salinity as Factors Affecting Spartina Alterniflora Growth and Dieback, Ph.D. dissertation, North Carolina State University, Dept. of Biology, Raleigh, N.C.

Little, R. J. A., and Rubin, D. B. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Mackie, J. L. 1974. *The Cement of the Universe*. Oxford, U.K.: Oxford University Press.

Madigan, D., and Raftery, A. 1994. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association* 89(428): 1535–1546.

Madigan, D.; Andersson, S. A.; Perlman, M. D.; and Volinsky, C. T. 1996. Bayesian Model Averaging and Model Selection for Markov Equivalence Classes of Acyclic Digraphs. *Communications in Statistics—Theory and Methods* 25(11): 2493–2519.

Manski, C. F. 1995. *Identification Problems in the Social Sciences*. Cambridge, Mass.: Harvard University Press.

Meek, C. 1998. Selecting Graphical Models: Causal and Statistical Modeling, Ph.D. dissertation, Department of Philosophy, Carnegie Mellon University.

Meek, C. 1995a. Causal Inference and Causal Explanation with Background Knowledge. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 403–410. San Francisco: Morgan Kaufmann Publishers.

Meek, C. 1995b. Strong Completeness and Faithfulness in Bayesian Networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 411–418. San Francisco: Morgan Kaufmann Publishers.

Meyer, B. D. 1995. Natural and Quasi-Experiments in Economics. *Journal of Business and Economic Statistics* 13(2): 151–161.

Neapolitan, R. 1990. *Probabilistic Reasoning in Expert Systems*. New York: Wiley.

Pearl, J. 1996. On the Foundation of Structural Equation Models, or When Can We Give Causal Interpretation to Structural Coefficients?, Technical Report, R-244-S, Cognitive Systems Laboratory, Department of Computer Science, University of California.

Pearl, J. 1995. Causal Diagrams for Empirical Research. *Biometrika* 82:669–709.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Francisco, Calif.: Morgan Kaufmann.

Pearl, J., and Verma, T. S. 1991. A Theory of Inferred Causality. In *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning*, 441–452. San Francisco: Morgan Kaufmann Publishers.

Pratt, J., and Schlaifer, R. 1988. On the Interpretation and Observation of Laws. *Journal of Econometrics* 39(1–2): 23–52.

Provan, G. M., and Clarke, J. R. 1993. Dynamic Network Construction and Updating Techniques for the Diagnosis of Acute Abdominal Pain. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(3): 299–306.

Reichenbach, H. 1956. *The Direction of Time*. Berkeley, Calif.: University of California Press.

Robins, J. 1989. The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies. In *Health Services Research Methodology: A Focus on AIDS*, eds. L. Sechrest, H. Freeman, and A. Mulley, 113–159. Washington, D.C.: U.S. Public Health Service.

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(2): 1393–1512.

Rubin, D. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66(5): 688–701.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, N.J.: Princeton University Press.

Savage, L. J. 1954. *Foundations of Statistics*. New York: Wiley.

Scheines, R.; Spirtes, P.; Glymour, C.; and Meek, C. 1995. *TETRAD II: Tools for Causal Modeling* (with software). Hillsdale, N.J.: Lawrence Erlbaum.

Simon, H. 1953. Causal Ordering and Identifiability. In *Studies in Econometric Method*, eds. W. C. Hood and T. C. Koopmans, 49–74. New York: Wiley.

Simpson, C. 1951. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society* B13:238–241.

Singh, M., and Valtorta, M. 1993. An Algorithm for the Construction of Bayesian Network Structures from Data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 259–265. San Francisco: Morgan Kaufmann Publishers.

Spirtes, P., and Meek, C. 1995. Learning Bayesian Networks with Discrete Variables from Data. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 294–299. Menlo Park, Calif.: AAAI Press.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. New York: Springer-Verlag. Available at hss.cmu.edu/html/ departments/philosophy/TETRAD.BOOK/book.html.

Spirtes, P.; Glymour, C.; and Scheines, R. 1991. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review* 9 (1): 62–72.

Spirtes, P.; Richardson, T.; and Meek, C. 1997. Heuristic Greedy Search Algorithms for Latent Variable Models. Paper presented at the InternationalWorkshop on Artificial Intelligence and Statistics, January 4–7, Ft. Lauderdale, Florida.

Suppes, P. 1970. *A Probabilistic Theory of Causality*. Amsterdam, The Netherlands: North Holland.

Verma, T. S., and Pearl, J. 1990. Equivalence and Synthesis of Causal Models. Paper presented at the Conference on Uncertainty in Artificial Intelligence, July 27–29, Cambridge, Mass.

Wedelin, D. 1993. Efficient Algorithms for Probabilistic Inference, Combinatorial Optimization, and the Discovery of Causal Structure from Data, Ph.D. dissertation, Department of Computer Science, Chalmers University of Technology, Sweden.

Wellman, M. P., and Henrion, M. 1993. Explaining "Explaining Away." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(3): 287–292.

Wermuth, N., and Lauritzen, S. L. 1990. On Substantive Research Hypotheses, Conditional Independence Graphs, and Graphical Chain Models. *Journal of the Royal Statistical Society* B 52(1): 21–72.

Wright, S. 1921. Correlation and Causation. *Journal of Agricultural Research* 20(1): 557– 585.

Xiang, Y.; Wong, S. K. M.; and Cercone, N. 1996. Critical Remarks on Single Link Search in Learning Belief Networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 564–571. San Francisco: Morgan Kaufmann Publishers.

Yao, Q., and Tritchler, D. 1996. Likelihood-Based Causal Inference. In *Learning from Data: Lecture Notes in Statistics*, eds. D. Fisher D. and H. Lenz, 35–44. New York: Springer- Verlag.

Zurek, W. 1991. Decoherence and the Transition from Quantum to Classical. *Physics Today* 44 (10):36–44.

PART ONE

Causation, Representation and Prediction

Directed graphs and associated parameters can encode probability distributions, but what makes these representations about *causal* relations is that they also contain information about how the influence of interventions or manipulations of some variables propagates to other variables.

The first of the two chapters in this section was written in 1991 but has not been previously published. The chapter introduces the basic ideas used to compute the propagation of influence by means of causal graphs, and relates the representation, assumptions and procedures to a formalism—the "Rubin framework"— sometimes used in statistics for similar purposes. This chapter led to published work on procedures for calculating the propagation of influence when the causal and probabilistic structure is only partially known.

The second chapter in this section, by Judea Pearl, offers a diagnosis of the many conceptual confusions about causal prediction in the literature of social statistics, and also offers a solution. The diagnosis is that there is no standard language, no formal notation, to distinguish conditioning on a variable from intervening to fix its value. Of course, a notation is only good if the distinctions it allows can be used to good purpose, and Pearl uses the notation to formulate rules for causal prediction, which are illustrated in a variety of clear and striking examples.