CHAPTER TWO

A Theory of Inferred Causation

I would rather discover one causal law
than be King of Persia.
Democritus (460–370 B.C.)

Preface

The possibility of learning causal relationships from raw data has been on philosophers' dream lists since the time of Hume (1711–1776). That possibility entered the realm of formal treatment and feasible computation in the mid-1980s, when the mathematical relationships between graphs and probabilistic dependencies came to light. The approach described herein is an outgrowth of Rebane and Pearl (1987) and Pearl (1988b, Chap. 8), which describes how causal relationships can be inferred from nontemporal statistical data if one makes certain assumptions about the underlying process of data generation (e.g., that it has a tree structure). The prospect of inferring causal relationships from weaker structural assumptions (e.g., general directed acyclic graphs) has motivated parallel research efforts at three universities: UCLA, Carnegie Mellon University (CMU), and Stanford. The UCLA and CMU teams pursued an approach based on searching the data for patterns of conditional independencies that reveal fragments of the underlying structure and then piecing those fragments together to form a coherent causal model (or a set of such models). The Stanford group pursued a Bayesian approach, where data are used to update prior probabilities assigned to candidate causal structures (Cooper and Herskovits 1991). The UCLA and CMU efforts have led to similar theories and almost identical discovery algorithms, which were implemented in the TETRAD II program (Spirtes et al. 1993). The Bayesian approach has since been pursued by a number of research teams (Singh and Valtorta 1995; Heckerman et al. 1994) and now serves as the basis for several graph-based learning methods (Jordan 1998). This chapter describes the approach pursued by Tom Verma and me in the period 1988–1992, and it briefly summarizes related extensions, refinements, and improvements that have been advanced by the CMU team and others. Some of the philosophical rationale behind this development, primarily the assumption of minimality, are implicit in the Bayesian approach as well (Section 2.9.1).

The basic idea of automating the discovery of causes – and the specific implementation of this idea in computer programs – came under fierce debate in a number of forums (Cartwright 1995a; Humphreys and Freedman 1996; Cartwright 1999; Korb and Wallace 1997; McKim and Turner 1997; Robins and Wasserman 1999). Selected aspects of this debate will be addressed in the discussion section at the end of this chapter (Section 2.9.1).

Acknowledging that statistical associations do not *logically* imply causation, this chapter asks whether weaker relationships exist between the two. In particular, we ask:

1. What clues prompt people to perceive causal relationships in uncontrolled observations?
2. What assumptions would allow us to infer causal models from these clues?
3. Would the models inferred tell us anything useful about the causal mechanisms that underly the observations?

In Section 2.2 we define the notions of causal models and causal structures and then describe the task of causal discovery as an inductive game that scientists play against Nature. In Section 2.3 we formalize the inductive game by introducing "minimal model" semantics – the semantical version of Occam's razor – and exemplify how, contrary to common folklore, causal relationships can be distinguished from spurious covariations following this standard norm of inductive reasoning. Section 2.4 identifies a condition, called *stability* (or *faithfulness*), under which effective algorithms exist that uncover structures of casual influences as defined here. One such algorithm (called IC), introduced in Section 2.5, uncovers the set of all causal models compatible with the data, assuming all variables are observed. Another algorithm (IC*), described in Section 2.6, is shown to uncover many (though not all) valid causal relationships when some variables are *not* observable. In Section 2.7 we extract from the IC* algorithm the essential conditions under which causal influences are inferred, and we offer these as independent definitions of genuine influences and spurious associations, with and without temporal information. Section 2.8 offers an explanation for the puzzling yet universal agreement between the temporal and statistical aspects of causation. Finally, Section 2.9 summarizes the claims made in this chapter, re-explicates the assumptions that lead to these claims, and offers new justifications of these assumption in light of ongoing debates.

2.1 INTRODUCTION – THE BASIC INTUITIONS

An autonomous intelligent system attempting to build a workable model of its environment cannot rely exclusively on preprogrammed causal knowledge; rather, it must be able to translate direct observations to cause-and-effect relationships. However, given that statistical analysis is driven by covariation, not causation, and assuming that the bulk of human knowledge derives from passive observations, we must still identify the clues that prompt people to perceive causal relationships in the data. We must also find a computational model that emulates this perception.

Temporal precedence is normally assumed to be essential for defining causation, and it is undoubtedly one of the most important clues that people use to distinguish causal from other types of associations. Accordingly, most theories of causation invoke an explicit requirement that a cause precede its effect in time (Reichenbach 1956; Good 1961; Suppes 1970; Shoham 1988). Yet temporal information alone cannot distinguish genuine causation from spurious associations caused by unknown factors – the barometer falls before it rains yet does not cause the rain. In fact, the statistical and philosophical literature has adamantly warned analysts that, unless one knows in advance all causally relevant factors or unless one can carefully manipulate some variables, no genuine causal inferences are possible (Fisher 1935; Skyrms 1980; Cliff 1983; Eells and Sober 1983; Holland 1986; Gardenfors 1988; Cartwright 1989).[1] Neither condition is realizable in normal learning environments, and the question remains how causal knowledge is ever acquired from experience.

The clues that we explore in this chapter come from certain patterns of statistical associations that are characteristic of causal organizations – patterns that, in fact, can be given meaningful interpretation only in terms of causal directionality. Consider, for example, the following *intransitive* pattern of dependencies among three events: A and B are dependent, B and C are dependent, yet A and C are independent. If you ask a person to supply an example of three such events, the example would invariably portray A and C as two independent causes and B as their common effect, namely, $A \rightarrow B \leftarrow C$. (In my favorite example, A and C are the outcomes of two fair coins, and B represents a bell that rings whenever either coin comes up heads.) Fitting this dependence pattern with a scenario in which B is the cause and A and C are the effects is mathematically feasible but very unnatural (the reader is encouraged to try this exercise).

Such thought experiments tell us that certain patterns of dependency, void of temporal information, are conceptually characteristic of certain causal directionalities and not others. Reichenbach (1956), who was the first to wonder about the origin of those patterns, suggested that they are characteristic of Nature, reflective of the second law of thermodynamics. Rebane and Pearl (1987) posed the question in reverse, and asked whether the distinctions among the dependencies associated with the three basic causal substructures: $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$, and $X \rightarrow Y \leftarrow Z$ can be used to uncover genuine causal influences in the underlying data-generating process. They quickly realized that the key to determining the direction of the causal relationship between X and Y lies in "the presence of a third variable Z that correlates with Y but not with X," as in the collider $X \rightarrow Y \leftarrow Z$, and developed an algorithm that recovers both the edges and directionalities in the class of causal graphs that they considered (i.e., a polytrees).

The investigation in this chapter formalizes these intuitions and extends the Rebane-Pearl recovery algorithm to general graphs, including graphs with unobserved variables.

2.2 THE CAUSAL DISCOVERY FRAMEWORK

We view the task of causal discovery as an induction game that scientists play against Nature. Nature possesses stable causal mechanisms that, on a detailed level of descriptions, are deterministic functional relationships between variables, some of which are unobservable. These mechanisms are organized in the form of an acyclic structure, which the scientist attempts to identify from the available observations.

[1] Some of the popular quotes are: "No causation without manipulation" (Holland 1986), "No causes in, no causes out" (Cartwright 1989), "No computer program can take account of variables that are not in the analysis" (Cliff 1983). My favorite: "Causation first, Manipulation second."

Definition 2.2.1. (Causal Structure)
A causal structure *of a set of variables V is a directed acyclic graph (DAG) in which each node corresponds to a distinct element of V, and each link represents a direct functional relationship among the corresponding variables.*

A causal structure serves as a blueprint for forming a "causal model" – a precise specification of *how* each variable is influenced by its parents in the DAG, as in the structural equation model of (1.40). Here we assume that Nature is at liberty to impose arbitrary functional relationships between each effect and its causes and then to perturb these relationships by introducing arbitrary (yet mutually independent) disturbances. These disturbances reflect "hidden" or unmeasurable conditions that Nature governs by some undisclosed probability function.

Definition 2.2.2 (Causal Model)
A causal model *is a pair $M = \langle D, \Theta_D \rangle$ consisting of a causal structure D and a set of parameters Θ_D compatible with D. The parameters Θ_D assign a function $x_i = f_i(pa_i, u_i)$ to each $X_i \in V$ and a probability measure $P(u_i)$ to each u_i, where PA_i are the parents of X_i in D and where each U_i is a random disturbance distributed according to $P(u_i)$, independently of all other u.*

As we have seen in Chapter 1 (Theorem 1.4.1), the assumption of independent disturbances renders the model *Markovian* in the sense that each variable is independent of all its nondescendants, conditional on its parents in D. The ubiquity of the Markov assumption in human discourse may be reflective of the granularity of the models we deem useful for understanding Nature. We can start in the deterministic extreme, where all variables are explicated in microscopic detail and where the Markov condition certainly holds. As we move up to macroscopic abstractions by aggregating variables and introducing probabilities to summarize omitted variables, we need to decide at what stage the abstraction has gone too far and where useful properties of causation are lost. Evidently, the Markov condition has been recognized by our ancestors (the authors of our causal thoughts) as a property worth protecting in this abstraction; correlations that are not explained by common causes are considered spurious, and models containing such correlations are considered incomplete. The Markov condition guides us in deciding when a set of parents PA_i is considered complete in the sense that it includes *all* the relevant immediate causes of variable X_i. It permits us to leave some of these causes out of PA_i (to be summarized by probabilities), but not if they also affect other variables modeled in the system. If a set PA_i in a model is too narrow, there will be disturbance terms that influence several variables simultaneously, and the Markov property will be lost. Such disturbances will be treated explicitly as "latent" variables (see Definition 2.3.2). Once we acknowledge the existence of latent variables and represent their existence explicitly as nodes in a graph, the Markov property is restored.

Once a causal model M is formed, it defines a joint probability distribution $P(M)$ over the variables in the system. This distribution reflects some features of the causal structure (e.g., each variable must be independent of its grandparents, given the values of its parents). Nature then permits the scientist to inspect a select subset $O \subseteq V$ of "observed" variables and to ask questions about $P_{[O]}$, the probability distribution over the observables, but it hides the underlying causal model as well as the causal structure. We investigate the feasibility of recovering the topology D of the DAG from features of the probability distribution $P_{[O]}$.[2]

2.3 MODEL PREFERENCE (OCCAM'S RAZOR)

In principle, since V is unknown, there is an unbounded number of models that would fit a given distribution, each invoking a different set of "hidden" variables and each connecting the observed variables through different causal relationships. Therefore, with no restriction on the type of models considered, the scientist is unable to make any meaningful assertions about the structure underlying the phenomena. For example, every probability distribution $P_{[O]}$ can be generated by a structure in which no observed variable is a cause of another but instead all variables are consequences of one latent common cause, U.[3] Likewise, assuming $V = O$ but lacking temporal information, the scientist can never rule out the possibility that the underlying structure is a complete, acyclic, and arbitrarily ordered graph – a structure that (with the right choice of parameters) can *mimic* the behavior of any model, regardless of the variable ordering. However, following standard norms of scientific induction, it is reasonable to rule out any theory for which we find a simpler, less elaborate theory that is equally consistent with the data (see Definition 2.3.5). Theories that survive this selection process are called *minimal.* With this notion, we can construct our (preliminary) definition of inferred causation as follows.

Definition 2.3.1 (Inferred Causation (Preliminary))
A variable X is said to have a causal influence *on a variable Y if a directed path from X to Y exists in every minimal structure consistent with the data.*

Here we equate a causal structure with a scientific theory, since both contain a set of free parameters that can be adjusted to fit the data. We regard Definition 2.3.1 as preliminary because it assumes that all variables are observed. The next few definitions generalize the concept of minimality to structures with unobserved variables.

Definition 2.3.2 (Latent Structure)
A latent structure *is a pair $L = \langle D, O\rangle$, where D is a causal structure over V and where $O \subseteq V$ is a set of observed variables.*

Definition 2.3.3 (Structure Preference)
One latent structure $L = \langle D, O\rangle$ is preferred *to another $L' = \langle D', O\rangle$ (written $L \preceq L'$) if and only if D' can mimic D over O – that is, if and only if for every Θ_D there exists a*

[2] This formulation invokes several idealizations of the actual task of scientific discovery. It assumes, for example, that the scientist obtains the distribution directly, rather than events sampled from the distribution. Additionally, we assume that the observed variables actually appear in the original causal model and are not some aggregate thereof. Aggregation might result in feedback loops, which we do not discuss in this chapter.

[3] This can be realized by letting U have as many states as O, assigning to U the prior distribution $P(u) = P(o(u))$ (where $o(u)$ is the cell of O corresponding to state u), and letting each observed variable O_i take on its corresponding value in $o(u)$.

$\Theta'_{D'}$ such that $P_{[O]}(\langle D', \Theta'_{D'}\rangle) = P_{[O]}(\langle D, \Theta_D\rangle)$. Two latent structures are equivalent, *written $L' \equiv L$, if and only if $L \preceq L'$ and $L \succeq L'$.*[4]

Note that the preference for simplicity imposed by Definition 2.3.3 is gauged by the expressive power of a structure, not by its syntactic description. For example, one latent structure L_1 may invoke many more parameters than L_2 and still be preferred if L_2 can accommodate a richer set of probability distributions over the observables. One reason scientists prefer simpler theories is that such theories are more constraining and thus more falsifiable; they provide the scientist with less opportunities to overfit the data "hindsightedly" and therefore command greater credibility if a fit is found (Popper 1959; Pearl 1978; Blumer et al. 1987).

We also note that the set of independencies entailed by a causal structure imposes limits on its expressive power, that is, its power to mimic other structures. Indeed, L_1 cannot be preferred to L_2 if there is even one observable dependency that is permitted by L_1 and forbidden by L_2. Thus, tests for preference and equivalence can sometimes be reduced to tests of induced dependencies, which in turn can be determined directly from the topology of the DAGs without ever concerning ourselves with the set of parameters. This is the case in the absence of hidden variables (see Theorem 1.2.8) but does not hold generally in all latent structures. Verma and Pearl (1990) showed that some latent structures impose numerical rather than independence constraints on the observed distribution (see, e.g., Section 8.4, equations (8.21)–(8.23)); this makes the task of verifying model preference complicated but does still permit us to extend the semantical definition of inferred causation (Definition 2.3.1) to latent structures.

Definition 2.3.4 (Minimality)
A latent structure L is minimal *with respect to a class $\mathcal{L}$ of latent structures if and only if there is no member of $\mathcal{L}$ that is strictly preferred to L – that is, if and only if for every $L' \in \mathcal{L}$ we have $L \equiv L'$ whenever $L' \preceq L$.*

Definition 2.3.5 (Consistency)
A latent structure $L = \langle D, O\rangle$ is consistent *with a distribution $\hat{P}$ over O if D can accommodate some model that generates $\hat{P}$ – that is, if there exists a parameterization Θ_D such that $P_{[O]}(\langle D, \Theta_D\rangle) = \hat{P}$.*

Clearly, a necessary (and sometimes sufficient) condition for L to be consistent with $\hat{P}$ is that L can account for all the dependencies embodied in $\hat{P}$

Definition 2.3.6 (Inferred Causation)
Given $\hat{P}$, a variable C has a causal influence *on variable E if and only if there exists a directed path from C to E in every minimal latent structure consistent with $\hat{P}$.*

We view this definition as normative because it is based on one of the least disputed norms of scientific investigation: Occam's razor in its semantical casting. However, as with any

[4] We use the succinct term "preferred to" to mean "preferred or equivalent to," a relation that has also been named "a submodel of."

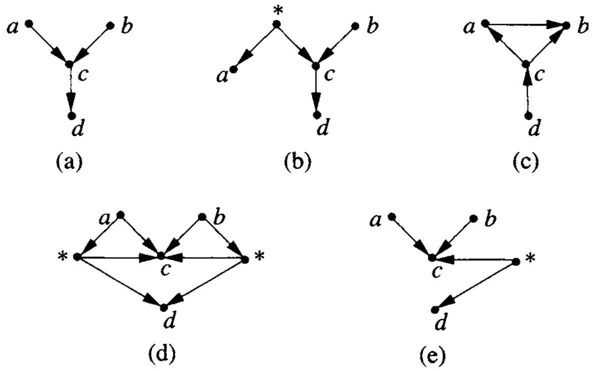

Figure 2.1 Causal structures illustrating the minimality of (a) and (b) and the justification for inferring the relationship $c \rightarrow d$. The node $(*)$ represents a hidden variable with any number of states.

scientific inquiry, we make no claims that this definition is guaranteed to always identify stable physical mechanisms in Nature. It identifies the mechanisms we can plausibly infer from nonexperimental data; moreover, it guarantees that any alternative mechanism will be less trustworthy than the one inferred because the alternative would require more contrived, hindsighted adjustment of parameters (i.e., functions) to fit the data.

As an example of a causal relation that is identified by Definition 2.3.6, imagine that observations taken over four variables $\{a, b, c, d\}$ reveal two independencies: "a is independent of b" and "d is independent of $\{a, b\}$ given c." Assume further that the data reveals *no other* independence besides those that logically follow from these two. This dependence pattern would be typical, for example, of the following variables: a = having a cold, b = having hay fever, c = having to sneeze, d = having to wipe one's nose. It is not hard to see that structures (a) and (b) in Figure 2.1 are minimal, for they entail the observed independencies and none other.[5] Furthermore, any structure that explains the observed dependence between c and d by an arrow from d to c, or by a hidden common cause $(*)$ between the two, cannot be minimal, because any such structure would be able to "out-mimic" the one shown in Figure 2.1(a) (or the one in Figure 2.1(b)), which reflects all observed independencies. For example, the structure of Figure 2.1(c), unlike that of Figure 2.1(a), accommodates distributions with arbitrary relations between a and b. Similarly, Figure 2.1(d) is not minimal because it fails to impose the conditional independence between d and $\{a, b\}$ given c and will therefore accommodate distributions in which d and $\{a, b\}$ are dependent given c. In contrast, Figure 2.1(e) is not consistent with the data, since it imposes an unobserved marginal independence between $\{a, b\}$ and d.

This example (taken from Pearl and Verma 1991) illustrates a remarkable connection between causality and probability: certain patterns of probabilistic dependencies (in our case, all dependencies except $(a \perp\!\!\!\perp b)$ and $(d \perp\!\!\!\perp \{a, b\} \mid c)$) imply unambiguous *causal* dependencies (in our case, $c \rightarrow d$) without making any assumption about the presence

[5] To verify that (a) and (b) are equivalent, we note that (b) can mimic (a) if we let the link $a \leftarrow *$ impose equality between the two variables. Conversely, (a) can mimic (b), since it is capable of generating every distribution that possesses the independencies entailed by (b). (For theory and methods of "reading off" conditional independencies from graphs, see Section 1.2.3 or Pearl 1988b.)

or absence of latent variables.[6] The only assumption invoked in this implication is minimality – models that overfit the data are ruled out.

2.4 STABLE DISTRIBUTIONS

Although the minimality principle is sufficient for forming a normative theory of inferred causation, it does not guarantee that the structure of the actual data-generating model would be minimal, or that the search through the vast space of minimal structures would be computationally practical. Some structures may admit peculiar parameterizations that would render them indistinguishable from many other minimal models that have totally disparate structures. For example, consider a binary variable C that takes the value 1 whenever the outcomes of two fair coins (A and B) are the same and takes the value 0 otherwise. In the trivariate distribution generated by this parameterization, each pair of variables is marginally independent yet is dependent conditional on the third variable. Such a dependence pattern may in fact be generated by three minimal causal structures, each depicting one of the variables as causally dependent on the other two, but there is no way to decide among the three. In order to rule out such "pathological" parameterizations, we impose a restriction on the distribution called *stability*, also known as DAG-isomorphism or perfect-mapness (Pearl 1988b, p. 128) and faithfulness (Spirtes et al. 1993). This restriction conveys the assumption that all the independencies embedded in P are stable; that is, they are entailed by the structure of the model D and hence remain invariant to any change in the parameters Θ_D. In our example, only the correct structure (namely, $A \rightarrow C \leftarrow B$) will retain its independence pattern in the face of changing parameterizations – say, when the coins become slightly biased.

Definition 2.4.1 (Stability)
Let $I(P)$ denote the set of all conditional independence relationships embodied in P. A causal model $M = \langle D, \Theta_D \rangle$ generates a stable *distribution if and only if $P(\langle D, \Theta_D \rangle)$ contains no extraneous independences – that is, if and only if $I(P(\langle D, \Theta_D \rangle)) \subseteq I(P(\langle D, \Theta'_D \rangle))$ for any set of parameters Θ'_D.*

The stability condition states that, as we vary the parameters from Θ to Θ', no independence in P can be destroyed; hence the name "stability." Succinctly, P is a stable distribution of M if it "maps" the structure D of M, that is, $(X \perp\!\!\!\perp Y \mid Z)_P \Longleftrightarrow (X \perp\!\!\!\perp Y \mid Z)_D$ for any three sets of variables X, Y, and Z (see Theorem 1.2.5).

The relationship between minimality and stability can be illustrated using the following analogy. Suppose we see a picture of a chair and that we need to decide between two theories as follows.

T_1: The object in the picture is a chair.

T_2: The object in the picture is either a chair or two chairs positioned such that one hides the other.

[6] Standard probabilistic definitions of causality (e.g., Suppes 1970; Eells 1991) invariably require knowledge of all relevant factors that may influence the observed variables (see Section 7.5.3).

Our preference for T_1 over T_2 can be justified on two principles, one based on minimality and the other on stability. The minimality principle argues that T_1 is preferred to T_2 because the set of scenes composed of single objects is a proper subset of scenes composed of two or fewer objects and, unless we have evidence to the contrary, we should prefer the more specific theory. The stability principle rules out T_2 a priori, arguing that it would be rather unlikely for two objects to align themselves so as to have one perfectly hide the other. Such an alignment would be *unstable* relative to slight changes in environmental conditions or viewing angle.

The analogy with independencies is clear. Some independencies are *structural,* that is, they would persist for every functional–distributional parameterization of the graph. Others are sensitive to the precise numerical values of the functions and distributions. For example, in the structure $Z \leftarrow X \rightarrow Y$, which stands for the relations

$$z = f_1(x, u_1), \quad y = f_2(x, u_2), \tag{2.1}$$

the variables Z and Y will be independent, conditional on X, for all functions f_1 and f_2. In contrast, if we add an arrow $Z \rightarrow Y$ to the structure and use a linear model

$$z = \gamma x + u_1, \quad y = \alpha x + \beta z + u_2, \tag{2.2}$$

with $\alpha = -\beta\gamma$, then Y and X will be independent. However, the independence between Y and X is unstable because it disappears as soon as the equality $\alpha = -\beta\gamma$ is violated. The stability assumption presumes that this type of independence is unlikely to occur in the data, that all independencies are structural.

To further illustrate the relations between stability and minimality, consider the causal structure depicted in Figure 2.1(c). The minimality principle rejects this structure on the ground that it fits a broader set of distributions than those fitted by structure (a). The stability principle rejects this structure on the ground that, in order to fit the data (specifically, the independence $(a \perp\!\!\!\perp b)$), the association produced by the arrow $a \rightarrow b$ must cancel precisely the one produced by the path $a \leftarrow c \rightarrow b$. Such precise cancelation cannot be stable, for it cannot be sustained for all functions connecting variables a, b, and c. In structure (a), by contrast, the independence $(a \perp\!\!\!\perp b)$ is stable.

2.5 RECOVERING DAG STRUCTURES

With the added assumption of stability, every distribution has a unique minimal causal structure (up to d-separation equivalence), as long as there are no hidden variables. This uniqueness follows from Theorem 1.2.8, which states that two causal structures are equivalent (i.e., they can mimic each other) if and only if they relay the same dependency information – namely, they have the same skeleton and same set of v-structures.

In the absence of unmeasured variables, the search for the minimal model then boils down to reconstructing the structure of a DAG D from queries about conditional independencies, assuming that those independencies reflect d-separation conditions in some undisclosed underlying DAG D_0. Naturally, since D_0 may have equivalent structures, the reconstructed DAG will not be unique, and the best we can do is to find a graphical representation for the equivalence class of D_0. Such graphical representation was introduced in Verma and Pearl (1990) under the name *pattern*. A pattern is a partially directed

DAG, in particular, a graph in which some edges are directed and some are nondirected. The directed edges represent arrows that are common to every member in the equivalence class of D_0, while the undirected edges represent ambivalence; they are directed one way in some equivalent structures and another way in others.

The following algorithm, introduced in Verma and Pearl (1990), takes as input a stable probability distribution $\hat{P}$ generated by some underlying DAG D_0 and outputs a pattern that represents the equivalence class of D_0.[7]

IC Algorithm **(Inductive Causation)**

Input: $\hat{P}$, a stable distribution on a set V of variables.

Output: a pattern $H(\hat{P})$ compatible with $\hat{P}$.

1. For each pair of variables a and b in V, search for a set S_{ab} such that $(a \perp\!\!\!\perp b \mid S_{ab})$ holds in $\hat{P}$ – in other words, a and b should be independent in $\hat{P}$, conditioned on S_{ab}. Construct an undirected graph G such that vertices a and b are connected with an edge if and only if no set S_{ab} can be found.
2. For each pair of nonadjacent variables a and b with a common neighbor c, check if $c \in S_{ab}$.
   If it is, then continue.
   If it is not, then add arrowheads pointing at c (i.e., $a \rightarrow c \leftarrow b$).
3. In the partially directed graph that results, orient as many of the undirected edges as possible subject to two conditions: (i) Any alternative orientation would yield a new v-structure; or (ii) Any alternative orientation would yield a directed cycle.

The IC algorithm leaves the details of steps 1 and 3 unspecified, and several refinements have been proposed for optimizing these two steps. Verma and Pearl (1990) noted that, in sparse graphs, the search can be trimmed substantially if commenced with the Markov network of $\hat{P}$, namely, the undirected graph formed by linking only pairs that are dependent conditionally on all other variables. In linear Gaussian models, the Markov network can be found in polynomial time, through matrix inversion, by assigning edges to pairs that correspond to the nonzero entries of the inverse covariance matrix. Spirtes and Glymour (1991) proposed a general systematic way of searching for the sets S_{ab} in step 1. Starting with sets S_{ab} of cardinality 0, then cardinality 1, and so on, edges are recursively removed from a complete graph as soon as separation is found. This refinement, called the PC algorithm (after its authors, Peter and Clark), enjoys polynomial time in graphs of finite degree because, at every stage, the search for a separating set S_{ab} can be limited to nodes that are adjacent to a and b.

Step 3 of the IC algorithm can be systematized in several ways. Verma and Pearl (1992) showed that, starting with any pattern, the following four rules are required for obtaining a maximally oriented pattern.

[7] The IC algorithm, as introduced in Verma and Pearl (1990), was designed to operate on latent structures. For clarity, we here present the algorithm in two separate parts, IC and IC*, with IC restricted to DAGs and IC* operating on latent structures.

R_1: Orient b — c into $b \rightarrow c$ whenever there is an arrow $a \rightarrow b$ such that a and c are nonadjacent.

R_2: Orient a — b into $a \rightarrow b$ whenever there is chain $a \rightarrow c \rightarrow b$.

R_3: Orient a — b into $a \rightarrow b$ whenever there are two chains a — $c \rightarrow b$ and a — $d \rightarrow b$ such that c and d are nonadjacent.

R_4: Orient a — b into $a \rightarrow b$ whenever there are two chains a — $c \rightarrow d$ and $c \rightarrow d \rightarrow b$ such that c and b are nonadjacent and a and d are adjacent.

Meek (1995) showed that these four rules are also sufficient, so that repeated application will eventually orient *all* arrows that are common to the equivalence class of D_0. Moreover, R_4 is not required if the starting orientation is limited to v-structures.

Another systematization is offered by an algorithm due to Dor and Tarsi (1992) that tests (in polynomial time) if a given partially oriented acyclic graph can be fully oriented without creating a new v-structure or a directed cycle. The test is based on recursively removing any vertex v that has the following two properties:

1. no edge is directed outward from v;
2. every neighbor of v that is connected to v through an undirected edge is also adjacent to all the other neighbors of v.

A partially oriented acyclic graph has an admissible extension in a DAG if and only if all its vertices can be removed in this fashion. Thus, to find the maximally oriented pattern, we can (i) separately try the two orientations, $a \rightarrow b$ and $a \leftarrow b$, for every undirected edge a — b, and (ii) test whether both orientations, or just one, have extensions. The set of uniquely orientable arrows constitutes the desired maximally oriented pattern. Additional refinements can be found in Chickering (1995), Andersson et al. (1997), and Moole (1997).

Latent structures, however, require special treatment, because the constraints that a latent structure imposes upon the distribution cannot be completely characterized by any set of conditional independence statements. Fortunately, certain sets of those independence constraints can be identified (Verma and Pearl 1990); this permits us to recover valid fragments of latent structures.

2.6 RECOVERING LATENT STRUCTURES

When Nature decides to "hide" some variables, the observed distribution $\hat{P}$ need no longer be stable relative to the observable set O. That is, we are no longer guaranteed that, among the minimal latent structures compatible with $\hat{P}$, there exists one that has a DAG structure. Fortunately, rather than having to search through this unbounded space of latent structures, the search can be confined to graphs with finite and well-defined structures. For every latent structure L, there is a dependency-equivalent latent structure (the projection) of L on O in which every unobserved node is a root node with exactly two observed children. We characterize this notion explicitly as follows.

Definition 2.6.1 (Projection)
A latent structure $L_{[O]} = \langle D_{[O]}, O \rangle$ *is a* projection *of another latent structure L if and only if:*

1. *every unobservable variable of* $D_{[O]}$ *is a parentless common cause of exactly two nonadjacent observable variables; and*
2. *for every stable distribution P generated by L, there exists a stable distribution* P' *generated by* $L_{[O]}$ *such that* $I(P_{[O]}) = I(P'_{[O]})$.

Theorem 2.6.2 (Verma 1993)
Any latent structure has at least one projection.

It is convenient to represent projections using a bidirectional graph with only the observed variables as vertices (i.e., leaving the hidden variables implicit). Each bidirected link in such a graph represents a common hidden cause of the variables corresponding to the link's endpoints.

Theorem 2.6.2 renders our definition of inferred causation (Definition 2.3.6) operational; it can be shown (Verma 1993) that the existence of a certain link in a distinguished projection of any minimal model of $\hat{P}$ must indicate the existence of a causal path in every minimal model of $\hat{P}$. Thus, our search reduces to finding the distinguished projection of any minimal model of $\hat{P}$ and identifying the appropriate links. Remarkably, these links can be identified by a simple variant of the IC algorithm, here called IC*, that takes a stable distribution $\hat{P}$ and returns a *marked* pattern, which is a partially directed acyclic graph that contains four types of edges:

1. a marked arrow $a \xrightarrow{*} b$, signifying a directed path from a to b in the underlying model;
2. an unmarked arrow $a \to b$, signifying either a directed path from a to b or a latent common cause $a \leftarrow L \to b$ in the underlying model;
3. a bidirected edge $a \longleftrightarrow b$, signifying a latent common cause $a \leftarrow L \to b$ in the underlying model; and
4. an undirected edge $a — b$, standing for either $a \leftarrow b$ or $a \to b$ or $a \leftarrow L \to b$ in the underlying model.[8]

***IC* Algorithm* (Inductive Causation with Latent Variables)**

Input: $\hat{P}$, a stable distribution (with respect to some latent structure).

Output: core $(\hat{P})$, a marked pattern.

1. For each pair of variables a and b, search for a set S_{ab} such that a and b are independent in $\hat{P}$, conditioned on S_{ab}.

[8] Spirtes et al. (1993) used $a \circ\!\!\to b$ to represent uncertainty about the arrowhead at node a. Several errors in the original proof of IC* were pointed out to us by Peter Spirtes and were corrected in Verma (1993). Alternative proofs of correctness, as well as refinements in the algorithm, are given in Spirtes et al. (1993).

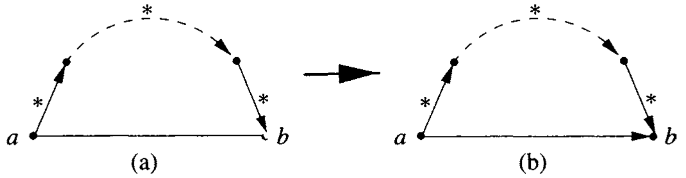

Figure 2.2 Illustration of R_2 in step 3 of the IC* algorithm.

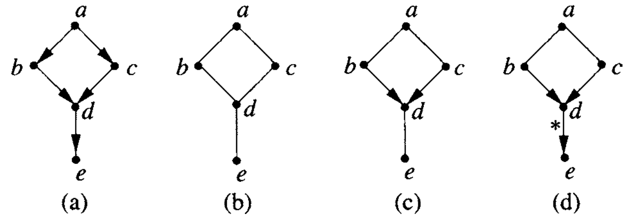

Figure 2.3 Graphs constructed by the IC* algorithm. (a) Underlying structure. (b) After step 1. (c) After step 2. (d) Output of IC*.

If there is no such S_{ab}, place an undirected link between the two variables, a — b.

2. For each pair of nonadjacent variables a and b with a common neighbor c, check if $c \in S_{ab}$.

   If it is, then continue.

   If it is not, then add arrowheads pointing at c (i.e., $a \rightarrow c \leftarrow b$).

3. In the partially directed graph that results, add (recursively) as many arrowheads as possible, and mark as many edges as possible, according to the following two rules:

   R_1: For each pair of nonadjacent nodes a and b with a common neighbor c, if the link between a and c has an arrowhead into c and if the link between c and b has no arrowhead into c, then add an arrowhead on the link between c and b pointing at b and mark that link to obtain $c \xrightarrow{*} b$.

   R_2: If a and b are adjacent and there is a directed path (composed strictly of marked links) from a to b (as in Figure 2.2), then add an arrowhead pointing toward b on the link between a and b.

Steps 1 and 2 of IC* are identical to those of IC, but the rules in step 3 are different; they do not orient edges but rather add arrowheads to the individual endpoints of the edges, thus accommodating bidirectional edges.

Figure 2.3 illustrates the operation of the IC* algorithm on the sprinkler example of Figure 1.2 (shown schematically in Figure 2.3(a)).

1. The conditional independencies entailed by this structure can be read off using the d-separation criterion (Definition 1.2.3), and the smallest conditioning sets corresponding to these independencies are given by $S_{ad} = \{b, c\}$, $S_{ae} = \{d\}$,

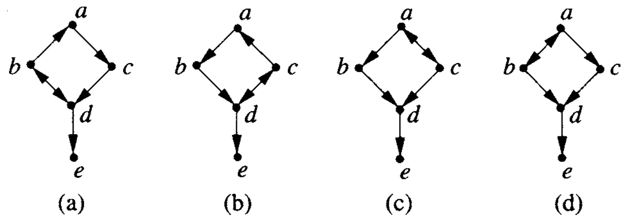

Figure 2.4 Latent structures equivalent to those of Figure 2.3(a).

$S_{bc} = \{a\}$, $S_{be} = \{d\}$, and $S_{ce} = \{d\}$. Thus, step 1 of IC* yields the undirected graph of Figure 2.3(b).

2. The triplet (b, d, c) is the only one that satisfies the condition of step 2, since d is not in S_{bc}. Accordingly, we obtain the partially directed graph of Figure 2.3(c).
3. Rule R_1 of step 3 is applicable to the triplet (b, d, e) (and to (c, d, e)), since b and e are nonadjacent and there is an arrowhead at d from b but not from e. We therefore add an arrowhead at e, and mark the link, to obtain Figure 2.3(d). This is also the final output of IC*, because R_1 and R_2 are no longer applicable.

The absence of arrowheads on a — b and a — c, and the absence of markings on $b \overset{*}{\rightarrow} d$ and $c \overset{*}{\rightarrow} d$, correctly represent the ambiguities presented by $\hat{P}$. Indeed, each of the latent structures shown in Figure 2.4 is observationally equivalent to that of Figure 2.3(a). Marking the link $d \overset{*}{\rightarrow} e$ in Figure 2.3(d) advertises the existence of a directed link $d \overset{*}{\rightarrow} e$ in each and every latent structure that is independence-equivalent to the one in Figure 2.3(a).

2.7 LOCAL CRITERIA FOR INFERRING CAUSAL RELATIONS

The IC* algorithm takes a distribution $\hat{P}$ and outputs a partially directed graph. Some of the links are marked unidirectional (denoting genuine causation), some are *un*marked unidirectional (denoting potential causation), some are bidirectional (denoting spurious association), and some are undirected (denoting relationships that remain undetermined). The conditions that give rise to these labelings can be taken as definitions for the various kinds of causal relationships. In this section we present explicit definitions of potential and genuine causation as they emerge from the IC* algorithm. Note that, in all these definitions, the criterion for causation between two variables (X and Y) will require that a third variable Z exhibit a specific pattern of dependency with X and Y. This is not surprising, since the essence of causal claims is to stipulate the behavior of X and Y under the influence of a third variable, one that corresponds to an external control of X (or Y) – as echoed in the paradigm of "no causation without manipulation" (Holland 1986). The difference is only that the variable Z, acting as a virtual control, must be identified within the data itself, as if Nature had performed the experiment. The IC* algorithm can be regarded as offering a systematic way of searching for variables Z that qualify as virtual controls, given the assumption of stability.

Definition 2.7.1 (Potential Cause)
A variable X has a potential causal influence *on another variable Y (that is inferable from $\hat{P}$) if the following conditions hold.*

1. *X* and *Y are dependent in every context.*
2. *There exists a variable Z and a context S such that*
   (i) *X and Z are independent given S (i.e., $X \perp\!\!\!\perp Z \mid S$) and*
   (ii) *Z and Y are dependent given S (i.e., $Z \not\!\perp\!\!\!\perp Y \mid S$).*

By "context" we mean a set of variables tied to specific values. In Figure 2.3(a), for example, variable b qualifies as a potential cause of d by virtue of variable $Z = c$ being dependent on d and independent of b in context $S = a$. Likewise, c qualifies as a potential cause of d (with $Z = b$ and $S = a$). Neither b nor c qualifies as a genuine cause of d, because this pattern of dependencies is also compatible with a latent common cause, shown as bidirected arcs in Figures 2.4(a)–(b). However, Definition 2.7.1 disqualifies d as a cause of b (or c), and this leads to the classification of d as a *genuine* cause of e, as formulated in Definition 2.7.2.[9] Note that Definition 2.7.1 precludes a variable X from being a potential cause of itself or of any other variable that functionally determines X.

Definition 2.7.2 (Genuine Cause)
A variable X has a genuine causal influence *on another variable Y if there exists a variable Z such that either:*

1. *X and Y are dependent in any context and there exists a context S satisfying*
   (i) *Z is a potential cause of X (per Definition 2.7.1),*
   (ii) *Z and Y are dependent given S (i.e., $Z \not\!\perp\!\!\!\perp Y \mid S$), and*
   (iii) *Z and Y are independent given $S \cup X$ (i.e., $Z \perp\!\!\!\perp Y \mid S \cup X$);*
   or
2. *X and Y are in the transitive closure of the relation defined in criterion 1.*

Conditions (i)–(iii) are illustrated in Figure 2.3(a) with $X = d$, $Y = e$, $Z = b$, and $S = \emptyset$. The destruction of the dependence between b and e through conditioning on d cannot be attributed to spurious association between d and e; genuine causal influence is the only explanation, as shown in the structures of Figure 2.4.

Definition 2.7.3 (Spurious Association)
Two variables X and Y are spuriously associated *if they are dependent in some context and there exist two other variables (Z_1 and Z_2) and two contexts (S_1 and S_2) such that:*

[9] Definition 2.7.1 was formulated in Pearl (1990) as a relation between events (rather than variables) with the added condition $P(Y \mid X) > P(Y)$ (in the spirit of Reichenbach 1956, Good 1961, and Suppes 1970). This refinement is applicable to any of the definitions in this section, but it will not be formulated explicitly.

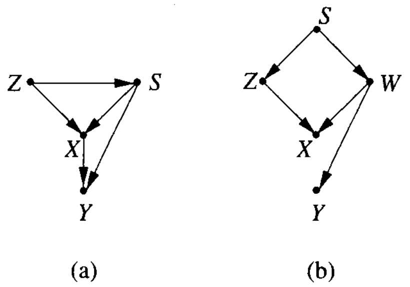

Figure 2.5 Illustration of how temporal information permits the inference of genuine causation and spurious associations (between X and Y) from the conditional independencies displayed in (a) and (b), respectively.

1. *Z_1 and X are dependent given S_1 (i.e., $Z_1 \not\perp\!\!\!\perp X \mid S_1$);*
2. *Z_1 and Y are independent given S_1 (i.e., $Z_1 \perp\!\!\!\perp Y \mid S_1$);*
3. *Z_2 and Y are dependent given S_2 (i.e., $Z_2 \not\perp\!\!\!\perp Y \mid S_2$); and*
4. *Z_2 and X are independent given S_2 (i.e., $Z_2 \perp\!\!\!\perp X \mid S_2$).*

Conditions 1 and 2 use Z_1 and S_1 to disqualify Y as a cause of X, paralleling conditions (i)–(ii) of Definition 2.7.1; conditions 3 and 4 use Z_2 and S_2 to disqualify X as a cause of Y. This leaves the existence of a latent common cause as the only explanation for the observed dependence between X and Y, as exemplified in the structure $Z_1 \rightarrow X \rightarrow Y \leftarrow Z_2$.

When temporal information is available (as is assumed in the most probabilistic theories of causality – Suppes 1970; Spohn 1983; Granger 1988), Definitions 2.7.2 and 2.7.3 simplify considerably because every variable preceding and adjacent to X now qualifies as a "potential cause" of X. Moreover, adjacency (i.e., condition 1 of Definition 2.7.1) is not required as long as the context S is confined to be earlier than X. These considerations lead to simpler conditions distinguishing genuine from spurious causes, as shown next.

Definition 2.7.4 (Genuine Causation with Temporal Information)
A variable X has a causal influence on Y if there is a third variable Z and a context S, both occurring before X, such that:

1. $(Z \not\perp\!\!\!\perp Y \mid S)$;
2. $(Z \perp\!\!\!\perp Y \mid S \cup X)$.

The intuition behind Definition 2.7.4 is the same as for Definition 2.7.2, except that temporal precedence is now used to establish Z as a potential cause of X. This is illustrated in Figure 2.5(a): If conditioning on X can turn Z and Y from dependent to independent (in context S), it must be that the dependence between Z and Y was mediated by X; given that Z precedes X, such mediation implies that X has a causal influence on Y.

Definition 2.7.5 (Spurious Association with Temporal Information)
Two variables X and Y are spuriously associated *if they are dependent in some context S, if X precedes Y, and if there exists a variable Z satisfying:*

1. $(Z \perp\!\!\!\perp Y \mid S)$;
2. $(Z \not\!\perp\!\!\!\perp X \mid S)$.

Figure 2.5(b) illustrates the intuition behind Definition 2.7.5. Here the dependence between X and Y cannot be attributed to causal connection between the two because such a connection would imply dependence between Z and Y, which is ruled out by condition 1.[10]

Examining the definitions just presented, we see that all causal relations are inferred from at least three variables. Specifically, the information that permits us to conclude that one variable is not a causal consequence of another comes in the form of an "intransitive triplet" – for example, the variables *a, b, c* in Figure 2.1(a) satisfying $(a \perp\!\!\!\perp b \mid \emptyset)$, $(a \not\!\perp\!\!\!\perp c \mid \emptyset)$, and $(b \not\!\perp\!\!\!\perp c \mid \emptyset)$. The argument goes as follows. If we find conditions (S_{ab}) where the variables a and b are each correlated with a third variable c but are independent of each other, then the third variable cannot act as a cause of a or b (recall that, in stable distributions, the presence of a common cause implies dependence among the effects); rather, c must either be their common effect $(a \rightarrow c \leftarrow b)$ or be associated with a and b via common causes, forming a pattern such as $a \quad \longleftrightarrow \quad c \quad \longleftrightarrow \quad b$. This is indeed the condition that permits the IC* algorithm to begin orienting edges in the graph (step 2) and to assign arrowheads pointing at c. It is also this intransitive pattern that is used to ensure that X is not a consequence of Y in Definition 2.7.1 and that Z is not a consequence of X in Definition 2.7.2. In Definition 2.7.3 we have two intransitive triplets, (Z_1, X, Y) and (X, Y, Z_2), thus ruling out direct causal influence between X and Y and so implying that spurious associations are the only explanation for their dependence.

This interpretation of intransitive triplets involves a virtual control of the effect variable, rather than of the putative cause; this is analogous to testing the null hypothesis in the manipulative view of causation (Section 1.3). For example, one of the reasons people insist that the rain causes the grass to become wet, and not the other way around, is that they can easily find other means of getting the grass wet that are totally independent of the rain. Transferred to our chain $a — c — b$, we preclude c from being a cause of a if we find another means (b) of potentially controlling c without affecting a (Pearl 1988a, p. 396). The analogy is merely heuristic, of course, because in observational studies we must wait for Nature to provide the appropriate control and refrain from contaminating that control with spurious associations (with a).

2.8 NONTEMPORAL CAUSATION AND STATISTICAL TIME

Determining the direction of causal influences from nontemporal data raises some interesting philosophical questions about the relationships between time and causal explanations. For example, can the orientation assigned to the arrow $X \rightarrow Y$ in Definitions 2.7.2

[10] Recall that transitivity of causal dependencies is implied by stability. Although it is possible to construct causal chains $Z \rightarrow X \rightarrow Y$ in which Z and Y are independent, such independence will not be sustained for *all* parameterizations of the chain.

or 2.7.4 ever clash with the available temporal information (say, by a subsequent discovery that Y precedes X)? Since the rationale behind Definition 2.7.4 is based on strong intuitions about the statistical aspects of causal relationships (e.g., no correlation without some causation), it is apparent that such clashes, if they occur, are rather rare. The question then arises: Why should orientations determined solely by statistical dependencies have anything to do with the flow of time?

In human discourse, causal explanations satisfy two expectations: temporal and statistical. The temporal aspect is represented by the understanding that a cause should precede its effect. The statistical aspect expects a complete causal explanation to screen off its various effects (i.e., render the effects conditionally independent);[11] explanations that do not screen off their effects are considered "incomplete," and the residual dependencies are considered "spurious" or "unexplained." The clashless coexistence of these two expectations through centuries of scientific observations implies that the statistics of natural phenomena must exhibit some basic temporal bias. Indeed, we often encounter phenomenon where knowledge of a present state renders the variables of the future state conditionally independent (e.g., multivariate economic time series as in (2.3)). However, we rarely find the converse phenomenon, where knowledge of the present state would render the components of the past state conditionally independent. Is there any compelling reason for this temporal bias?

A convenient way to formulate this bias is through the notion of statistical time.

Definition 2.8.1 (Statistical Time)
Given an empirical distribution P, a statistical time *of P is any ordering of the variables that agrees with at least one minimal causal structure consistent with P.*

We see, for example, that a scalar Markov chain process has many statistical times; one coinciding with the physical time, one opposite to it, and others that correspond to orderings that agree with any orientation of the Markov chain away from one of the nodes (arbitrarily chosen as a root). On the other hand, a process governed by two coupled Markov chains, such as

$$\begin{aligned} X_t &= \alpha X_{t-1} + \beta Y_{t-1} + \xi_t, \\ Y_t &= \gamma X_{t-1} + \delta Y_{t-1} + \eta_t, \end{aligned} \tag{2.3}$$

has only one statistical time – the one coinciding with the physical time.[12] Indeed, running the IC algorithm on samples taken from such a process – while suppressing all temporal information – quickly identifies the components of X_{t-1} and Y_{t-1} as genuine

[11] This expectation, known as Reichenbach's "conjunctive fork" or "common-cause" criterion (Reichenbach 1956; Suppes and Zaniotti 1981; Sober and Barrett 1992) has been criticized by Salmon (1984a), who showed that some events qualify as causal explanations though they fail to meet Reichenbach's criterion. However, Salmon's examples involve incomplete explanations, as they leave out variables that mediate between the cause and its various effects (see Section 2.9.1).

[12] Here ξ_t and η_t, are assumed to be two independent, white-noise time series. Also, $\alpha \neq \delta$ and $\gamma \neq \beta$.

causes of X_t and Y_t. This can be seen from Definition 2.7.1 (where X_{t-2} qualifies as a potential cause of X_{t-1} using $Z = Y_{t-2}$ and $S = \{X_{t-3}, Y_{t-3}\}$) and Definition 2.7.2 (where X_{t-1} qualifies as a genuine cause of X_t using $Z = X_{t-2}$ and $S = \{Y_{t-1}\}$).

The temporal bias postulated earlier can be expressed as follows.

Conjecture 2.8.2 (Temporal Bias)
In most natural phenomenon, the physical time coincides with at least one statistical time.

Reichenbach (1956) attributed the asymmetry associated with his conjunctive fork to the second law of thermodynamics. It is doubtful that the second law can provide a full account of the temporal bias just described, since the influence of the external noise ξ_t and η_t renders the process in (2.3) nonconservative.[13] Moreover, the temporal bias is language-dependent. For example, expressing (2.3) in a different coordinate system – say, using a linear transformation

$$X'_t = aX_t + bY_t,$$
$$Y'_t = cX_t + dY_t$$

– it is possible to make the statistical time in the (X', Y') representation run contrary to the physical time; that is, X'_t and Y'_t will be independent of each other conditional on their future values (X'_{t+1} and Y'_{t+1}) rather than their past values. This suggests that the consistent agreement between physical and statistical times is a by-product of the human choice of linguistic primitives and not a feature of physical reality. For example, if X_t and Y_t stand for the positions of two interacting particles at time t, with X'_t the position of their center of gravity and Y'_t their relative distance, then describing the particles' motion in the (X, Y) versus (X', Y') coordinate system is (in principle) a matter of choice. Evidently, however, this choice is not entirely whimsical; it reflects a preference toward coordinate systems in which the *forward* disturbances (ξ_t and η_t in (2.3)) are orthogonal to each other, rather than the corresponding backward disturbances (ξ'_t and η'_t). Pearl and Verma (1991) speculated that this preference represents survival pressure to facilitate predictions of future events, and that evolution has evidently ranked this facility more urgent than that of finding hindsighted explanations for current events. Whether this or some other force has shaped our choice of language remains to be investigated (see discussions in Price 1996), which makes the statistical–temporal agreement that much more interesting.

2.9 CONCLUSIONS

The theory presented in this chapter shows that, although statistical analysis cannot distinguish genuine causation from spurious covariation in every conceivable case, in many cases it can. Under the assumptions of model minimality (and/or stability), there are

[13] I am grateful to Seth Lloyd for this observation.

patterns of dependencies that should be sufficient to uncover genuine causal relationships. These relationships cannot be attributed to hidden causes lest we violate one of the basic maxims of scientific methodology: the semantical version of Occam's razor. Adherence to this maxim may explain why humans reach consensus regarding the directionality and nonspuriousness of causal relationships in the face of opposing alternatives that are perfectly consistent with observation. Echoing Cartwright (1989), we summarize our claim with the slogan "No causes in – No causes out; Occam's razor in – Some causes out."

How safe are the causal relationships inferred by the IC algorithm – or by the TETRAD program of Spirtes et al. (1993) or the Bayesian methods of Cooper and Herskovits (1991) or Heckerman et al. (1994)?

Recasting this question in the context of visual perception, we may equally well ask: How safe are our predictions when we recognize three-dimensional objects from their two-dimensional shadows, or from the two-dimensional pictures that objects reflect on our retinas? The answer is: Not absolutely safe, but good enough to tell a tree from a house and good enough to make useful inferences without having to touch every physical object that we see. Returning to causal inference, our question then amounts to assessing whether there are enough discriminating clues in a typical learning environment (say, in skill acquisition tasks or in epidemiological studies) to allow us to make reliable discriminations between cause and effect. Rephrased as a logical guarantee, we can categorically assert that the IC* algorithm will never label an arrow $a \rightarrow b$ as genuine if in fact a has no causal influence on b and if the observed distribution is stable relative to its underlying causal model.

On the practical side, we have shown that the assumption of model minimality, together with that of "stability" (no accidental independencies) lead to an effective algorithm for structuring candidate causal models capable of generating the data, transparent as well as latent. Simulation studies conducted at our laboratory in 1990 showed that networks containing tens of variables require fewer than 5,000 samples to have their structure recovered by the algorithm. For example, 1,000 samples taken from (a binary version of) the process shown in (2.3), each containing ten successive X, Y pairs, were sufficient to recover its double-chain structure (and the correct direction of time). The greater the noise, the quicker the recovery (up to a point). In testing this modeling scheme on real-life data, we have examined the observations reported in Sewal Wright's seminal paper "Corn and Hog Correlations" (Wright 1925). As expected, corn price (X) can clearly be identified as a cause of hog price (Y) but not the other way around. The reason lies in the existence of the variable corn crop (Z), which satisfies the conditions of Definition 2.7.2 (with $S = \emptyset$). Several applications of the principles and algorithms discussed in this chapter are described in Glymour and Cooper (1999, pp. 441–541).

It should be natural to ask how the new criteria for causation could benefit current research in machine learning and data mining. In some sense, our method resembles a standard, machine-learning search through a space of hypotheses (Mitchell 1982) where each hypothesis stands for a causal model. Unfortunately, this is where the resemblance ends. The prevailing paradigm in the machine-learning literature has been to define each hypothesis (or theory, or concept) as a subset of observable instances; once we observe the entire extension of this subset, the hypothesis is defined unambiguously. This is not the case in causal discovery. Even if the training sample exhausts the hypothesis subset (in our case, this corresponds to observing P precisely), we are still left with a vast number of equivalent causal theories, each stipulating a drastically different set of causal claims. Therefore, *fitness to data is an insufficient criterion for validating causal theories*. Whereas in traditional learning tasks we attempt to generalize from one set of instances to another, the causal modeling task is to generalize from behavior under one set of conditions to behavior under another set. Causal models should therefore be chosen by a criterion that challenges their stability against changing conditions, and this is indeed what scientists attempt to accomplish through controlled experimentation. Absent such experimentation, the best one can do is to rely on virtual control variables, like those revealed by Nature through the dependence patterns of Definitions 2.7.1–2.7.4.

2.9.1 On Minimality, Markov, and Stability

The idea of inferring causation from association cannot be expected to go unchallenged by scientists trained along the lines of traditional doctrines. Naturally, the assumptions underlying the theory described in this chapter – minimality and stability – come under attack from statisticians and philosophers. This section contains additional thoughts in defense of these assumptions.

Although few have challenged the principle of minimality (to do so would amount to challenging scientific induction), objections have been voiced against the way we defined the objects of minimization – namely, causal models. Definition 2.2.2 assumes that the stochastic terms u_i are mutually independent, an assumption that endows each model with the Markov property: conditioned on its parents (direct causes), each variable is independent of its nondescendants. This implies, among the other ramifications of d-separation, several familiar relationships between causation and association that are usually associated with Reichenbach's (1956) principle of common cause – for example, "no correlation without causation," "causes screen off their effects," "no action at a distance."

The Markovian assumption, as explained in our discussion of Definition 2.2.2, is a matter of convention, to distinguish complete from incomplete models.[14] By building the Markovian assumption into the definition of complete causal models (Definition 2.2.2) and then relaxing the assumption through latent structures (Definition 2.3.2), we declare our preparedness to miss the discovery of non-Markovian causal models that cannot be described as latent structures. I do not consider this loss to be very serious, because such models – even if any exist in the macroscopic world – would have limited utility as guides to decisions. For example, it is not clear how one would predict the effects of interventions from such a model, save for explicitly listing the effect of every conceivable intervention in advance.

[14] Discovery algorithms for certain non-Markovian models, involving cycles and selection bias, have been reported in Spirtes et al. (1995) and Richardson (1996).

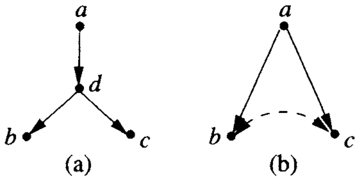

Figure 2.6 (a) Interactive fork. (b) Latent structure equivalent to (a).

It is not surprising, therefore, that criticisms of the Markov assumption, most notably those of Cartwright (1995a, 1997) and Lemmer (1993), have two characteristics in common:

1. they present macroscopic non-Markovian counterexamples that are reducible to Markovian latent structures of the type considered by Salmon (1984), that is, interactive forks; and
2. they propose no alternative, non-Markovian models from which one could predict the effects of actions and action combinations.

The interactive fork model is shown in Figure 2.6(a). If the intermediate node d is unobserved (or unnamed), then one is tempted to conclude that the Markov assumption is violated, since the observed cause (a) does not screen off its effects (b and c). The latent structure of Figure 2.6(b) can emulate the one of Figure 2.6(a) in all respects; the two can be indistinguishable both observationally and experimentally.

Only quantum-mechanical phenomena exhibit associations that cannot be attributed to latent variables, and it would be considered a scientific miracle if anyone were to discover such peculiar associations in the macroscopic world. Still, critics of the Markov condition insist that certain alleged counterexamples must be modeled as $P(bc \mid a)$ and not as $\Sigma_d P(b \mid d, a)P(c \mid d, a)$ – assuming, perhaps, that some insight or generality would be gained by leaving the dependency between b and c unexplained.

Ironically, perhaps the strongest evidence for the ubiquity of the Markov condition can be found in the philosophical program known as "probabilistic causality" (see Section 7.5), of which Cartwright is a leading proponent. In this program, causal dependence is defined as a probabilistic dependence that persists after conditioning on some set of relevant factors (Good 1961; Suppes 1970; Skyrms 1980; Cartwright 1983; Eells 1991). This definition rests on the assumption that conditioning on the right set of factors enables one to suppress all spurious associations – an assumption equivalent to the Markov condition. The intellectual survival of probabilistic causality as an active philosophical program for the past 40 years attests to the fact that counterexamples to the Markov condition are relatively rare and can be explained away through latent variables.

I now address the assumption of stability. The argument usually advanced to justify stability (Spirtes et al. 1993) appeals to the fact that strict equalities among products of parameters have zero Lebesgue measure in any probability space in which parameters can vary independently of one another. For example, the equality $\alpha = -\beta\gamma$ in the model of (2.2) has zero probability if we consider any continuous joint density over the parameters α, $-\beta$, and γ, unless that density somehow embodies the constraint

$\alpha = -\beta\gamma$ on a priori grounds. Freedman (1997), in contrast, claimed that there is no reason to assume that parameters are not in fact tied together by equality constraints of this sort, which would render the resulting distribution unstable (using Definition 2.4.1).

Indeed, the conditional independencies that a causal model advertises amount to none other than equality constraints on the joint distribution. The chain model $Y \rightarrow X \rightarrow Z$, for example, entails the equality

$$\rho_{YZ} = \rho_{XZ} \cdot \rho_{YX},$$

where ρ_{XY} is the correlation coefficient between X and Y; this equality constraint ties the three correlation coefficients in a permanent bond. What, then, gives equalities among correlation coefficients a privileged status over equalities among another set of parameters – say, α, β, and γ? Why do we consider the equality $\rho_{YZ} = \rho_{XZ} \cdot \rho_{YX}$ "stable" and the equality $\alpha = -\beta\gamma$ "accidental"?

The answer, of course, rests on the notion of *autonomy* (Aldrich 1989), a notion at the heart of all causal concepts (see Sections 1.3 and 1.4). A causal model is not just another scheme of encoding probability distribution through a set of parameters. The distinctive feature of causal models is that each variable is determined by a set of other variables through a relationship (called "mechanism") that remains *invariant* when other mechanisms are subjected to external influences. This invariance means that mechanisms *can* vary independently of one another, which in turns implies that the set of structural coefficients (e.g., α, β, γ in our example of (2.2)) – rather than other types of parameters (e.g., ρ_{YZ}, ρ_{XZ}, ρ_{YX}) – can and will vary independently when experimental conditions change. Consequently, equality constraints of the form $\alpha = -\beta\gamma$ are contrary to the idea of autonomy and will rarely occur under natural conditions. It is people's quest for stability that explains why they find it impossible to exemplify intransitive patterns of dependencies except by envisioning a common effect of two independent causes (see page 43). Any story that convincingly exemplifies a given pattern of dependencies must sustain that pattern regardless of the numerical values of the story parameters – a requirement we called "stability."

For this reason, it has been suggested that causal discovery methods based solely on associations, like those embodied in the IC* algorithm or the TETRAD-II program, will find their greatest potential in longitudinal studies conducted under slightly varying conditions, where accidental independencies are destroyed and only structural independencies are preserved. This assumes that, under such varying conditions, the parameters of the model will be perturbed while its structure remains intact – a delicate balance that might be hard to verify. Still, considering the alternative of depending only on controlled, randomized experiments, such longitudinal studies are an exciting opportunity.

Relation to the Bayesian Approach

It is important to stress that elements of the principles of minimality and stability also underlie causal discovery in the Bayesian approach. In this approach, one assigns prior probabilities to a set of candidate causal networks, based on their structures and parameters, and then uses Bayes's rule to score the degree to which a given network fits the data (Cooper and Herskovits 1991; Heckerman et al. 1999). A search is then conducted over the space of possible structures to seek the one(s) with the highest posterior score. Methods based on this approach have the advantage of operating well under small-sample conditions, but they encounter difficulties in coping with hidden variables. The assumption of parameter independence, which is made in all practical implementations of the Bayesian approach, induces preferences toward models with fewer parameters and hence toward minimality. Likewise, parameter independence can be justified only when the parameters represent mechanisms that are free to change independently of one another – that is, when the system is autonomous and hence stable.

Postscript for the Second Edition

Work on causal discovery has been pursued vigorously by the TETRAD group at Carengie Mellon University and reported in Spirtes et al. (2000), Robins et al. (2003), Scheines (2002), and Moneta and Spirtes (2006). Spirtes, Glymour, Scheines, and Tillman (2010) summarize the current state of the art in causal discovery.

Applications of causal discovery in economics are reported in Bessler (2002), Swanson and Granger (1997), and Demiralp and Hoover (2003). Gopnik et al. (2004) applied causal Bayesian networks to explain how children acquire causal knowledge from observations and actions (see also Glymour 2001).

Hoyer et al. (2006) and Shimizu et al. (2005, 2006) have proposed a new scheme of discovering causal directionality, based not on conditional independence but on functional composition. The idea is that in a linear model $X \rightarrow Y$ with non-Gaussian noise, variable Y is a linear combination of two independent noise terms. As a consequence, $P(y)$ is a convolution of two non-Gaussian distributions and would be, figuratively speaking, “more Gaussian” than $P(x)$. The relation of “more Gaussian than” can be given precise numerical measure and used to infer directionality of certain arrows.

Tian and Pearl (2001a,b) developed yet another method of causal discovery based on the detection of “shocks,” or spontaneous local changes in the environment which act like “Nature’s interventions,” and unveil causal directionality toward the consequences of those shocks.

Verma and Pearl (1990) noted that two latent structures may entail the same set of conditional independencies and yet impose different equality constraints on the joint distributions. These constraints, dubbed “dormant independencies,” were characterized systematically in Tian and Pearl (2002b) and Shpitser and Pearl (2008); they promise to provide a powerful new discovery tool for structure learning.

A program of benchmarks of causal discovery algorithms, named “Causality Workbench,” has been reported by Guyon et al. (2008a,b; http://clopinet.com/causality). Regular contests are organized in which participants are given real data or data generated by a concealed causal model, and the challenge is to predict the outcome of a select set of interventions.