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 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 experience. 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. This can only be determined by experiments—once we understand the logic behind the available clues and once we learn to piece these clues together coherently in large programs that tackle real-life problems.

The model-theoretic semantics presented in this chapter provides a conceptual and theoretical basis for such experiments. The IC* algorithm and the algorithms developed by the TETRAD group (Spirtes et al. 1993) demonstrate the computational feasibility of the approach. Waldmann et al. (1995) described psychological experiments on how

humans use the causal clues discussed in this chapter.

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 interesting to explore 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 modeling. 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 these show up in the data in the form of virtual control variables. Thus, the dependence patterns identified by Definitions 2.7.1–2.7.4 constitute islands of stability as well as virtual validation tests for causal models. It would be interesting to examine whether these criteria, when incorporated into existing machine-learning and data-mining programs, would improve the stability of relationships discovered by such programs.

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, and it has been adopted here as a useful abstraction of the underlying physical processes because such processes are too detailed to be of practical use. After all, investigators are free to decide what level of abstraction is useful for a given purpose, and Markovian models have been selected as targets of pursuit because

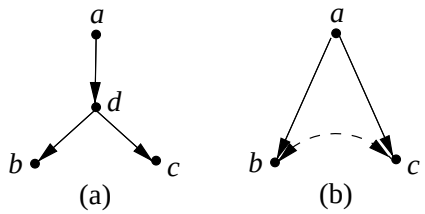

Figure 2.6: (a) Interactive fork. (b) Latent structure equivalent to (a).

of their usefulness in both prediction and decision making.[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 confess 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.

It is not surprising, therefore, that criticsism 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

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

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 $P(bc|a)$ and not as $\sum_d P(b|d,a)P(c|d,a)$—assuming, perhaps, that some insight or generality would be gained by leaving the dependency between b and c unexplained. The former model, in addition to being observationally indistinguishable from the latter, also leaves the causal effect $P_{ac}(b)$ unspecified. In contrast, the latent model predicts $P_{ac}(b) = P_a(b)$ and thus fulfills its role as a predictor of (experimentally testable) causal effects.

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 30 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, β, 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 constraints of this sort, which would render

the resulting distribution unstable (using Definition 2.4.1).

Freedman's critique receives unexpected support from the practice of structural modeling itself, where equality constraints are commonplace. 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 $\rho_{YZ} = \rho_{XZ} \cdot \rho_{YX}$ "substantive" and the equality $\alpha = -\beta\gamma$ "accidental," and why do we tie the notion of stability to the absence of the latter, not the former?

The answer, I believe, rests again 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. When we come to define mathematical objects such as causal models, we must ensure that the definition captures the distinct ways in which these objects are being used and conceptualized. 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 those other variables are subjected to external influences. Only by virtue of this invariance do causal models allow us to predict the effect of changes and interventions, capitalizing on the locality of such changes. 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 thus should not be considered part of the model.

For this reason, it has been suggested that causal modeling 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.