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.

---

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 $o(u)$.

Definition 2.3.2 (Latent Structure)
A latent structure *is a pair* $L = <D, O>$*, 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 $L = <D, O>$ *is* preferred *to another* $L' = <D', O>$ *(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* $\Theta'_{D'}$ *such that* $P_{[O]}(<D', \Theta'_{D'}>) = P_{[O]}(<D, \Theta_D>)$*. 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, i.e., 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

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

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 $L = <D, O>$ 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 $P_{[O]}(<D, \Theta_D>) = \hat{P}$.*

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 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

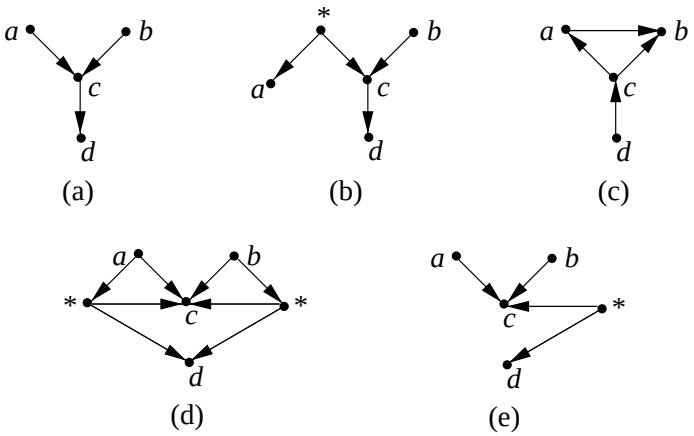

Figure 2.1: Causal structures illustrating the minimality of (a) and (b) and the justification for inferring the relationship $c \rightarrow d$. Asterics ($*$) represent hidden variables with any number of states.

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

[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].)

$\{a, b\}$ 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 $(d \perp\!\!\!\perp \{a, b\}|c)$) imply unambiguous *causal* dependencies (in our case, $c \rightarrow d$) without making any assumption about the presence or absence of latent variables.[6] The only assumption invoked in this implication is minimality—models that overfit the data are ruled out.

[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).