

1.4 Functional Causal Models

The way we have introduced the causal interpretation of Bayesian networks represents a fundamental departure from the way causal models (and causal graphs) were first introduced into genetics (Wright 1921), econometrics (Haavelmo 1943), and the social sciences (Duncan 1975), as well as from the way causal models are used routinely in physics and engineering. In those models, causal relationships are expressed in the form of deterministic, *functional* equations, and probabilities are introduced through the assumption that certain variables in the equations are unobserved. This reflects Laplace’s (1814) conception of natural phenomena, according to which nature’s laws are deterministic and randomness surfaces owing merely to our ignorance of the underlying boundary conditions. In contrast, all relationships in the definition of causal Bayesian networks were assumed to be inherently stochastic and thus appeal to the modern (i.e., quantum mechanical) conception of physics, according to which all nature’s laws are inherently probabilistic and determinism is but a convenient approximation.

In this book, we shall express preference toward Laplace’s quasi-deterministic conception of causality and will use it, often contrasted with the stochastic conception, to define and analyze most of the causal entities that we study. This preference is based on three considerations. First, the Laplacian conception is more general. Every stochastic model can be emulated by many functional relationships (with stochastic inputs), but not the other way around; functional relationships can only be approximated, as a limiting case, using stochastic models. Second, the Laplacian conception is more in tune with human intuition. The few esoteric quantum mechanical experiments that conflict with the predictions of the Laplacian conception evoke surprise and disbelief, and they demand that physicists give up deeply entrenched intuitions about locality and causality (Maudlin 1994). Our objective is to preserve, explicate, and satisfy—not destroy—those intuitions.¹²

¹²The often heard argument that human intuitions belong in psychology and not in science or philosophy is inapplicable when it comes to causal intuition—the original authors of causal thoughts cannot be ignored when the meaning of the concept is in question. Indeed, compliance with human intuition has been the ultimate criterion of adequacy in every philosophical study of causation, and

Finally, certain concepts that are ubiquitous in human discourse can be defined only in the Laplacian framework. We shall see, for example, that such simple concepts as “the probability that event B occurred *because* of event A ” and “the probability that event B would have been *different* if it were not for event A ” cannot be defined in terms of purely stochastic models. These so-called *counterfactual* concepts will require a synthesis of the deterministic and probabilistic components embodied in the Laplacian model.

1.4.1 Structural Equations

In its general form, a functional causal model consists of a set of equations of the form

$$x_i = f_i(pa_i, u_i), \quad i = 1, \dots, n, \quad (1.40)$$

where pa_i (connoting *parents*) stands for the set of variables judged to be immediate causes of X_i and where the U_i represent errors (or “disturbances”) due to omitted factors. Equation (1.40) is a nonlinear, nonparametric generalization of the linear structural equation models (SEMs)

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \quad i = 1, \dots, n, \quad (1.41)$$

which have become a standard tool in economics and social science (see Chapter 5 for a detailed exposition of this enterprise). In linear models, pa_i corresponds to those variables on the r.h.s. of (1.41) that have nonzero coefficients.

A set of equations in the form of (1.40) and in which each equation represents an autonomous mechanism is called *structural model*; if each mechanism determines the value of just one distinct variable (called the *dependent* variable), then the model is called a *structural causal model* or a *causal model* for short.¹³ Mathematically, the distinction between structural and algebraic equations is that the latter are characterized by

the proper incorporation of background information into statistical studies likewise relies on accurate interpretation of causal judgment.

¹³Formal treatment of causal models, structural equations, and error terms are given in Chapter 5 (Section 5.4.1) and Chapter 7 (Sections 7.1 and 7.2.5).

the set of solutions to the entire system of equations, whereas the former are characterized by the solutions of each individual equation. The implication is that any subset of structural equations is, in itself, a valid model of reality—one that prevails under some set of interventions.

To illustrate, Figure 1.5 depicts a canonical econometric model relating price and demand through the equations

$$q = b_1 p + d_1 i + u_1, \quad (1.42)$$

$$p = b_2 q + d_2 w + u_2, \quad (1.43)$$

where Q is the quantity of household demand for a product A , P is the unit price of product A , I is household income, W is the wage rate for producing product A , and u_1 and u_2 represent error terms—unmodeled factors that affect quantity and price, respectively (Goldberger 1992). The graph associated with this model is cyclic, and the vertices asso-

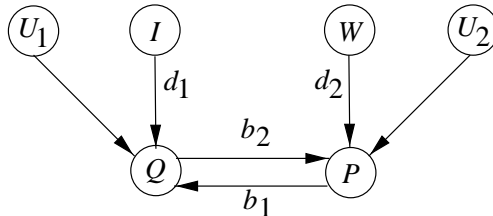


Figure 1.5: Causal diagram illustrating the relationship between price (P), demand (Q), income (Z), and wages (W).

ciated with the variables U_1 , U_2 , I , and W are root nodes, conveying the assumption of mutual independence. The idea of *autonomy* (Aldrich 1989), in this context, means that the two equations represent two loosely coupled segments of the economy, consumers and producers. Equation (1.42) describes how consumers decide what quantity Q to buy, and (1.43) describes how manufacturers decide what price P to charge. Like all feedback systems, this too represents implicit dynamics; today's prices are determined on the basis of yesterday's demand, and these prices will determine the demand in the next period of transactions. The solution to such equations represents a long-term

equilibrium under the assumption that the background quantities, U_1 and U_2 , remain constant.

The two equations are considered to be “autonomous” relative to the dynamics of changes in the sense that external changes affecting one equation do not imply changes to the others. For example, if government decides on price control and sets the price P at p_0 , then (1.43) will be modified to read $p = p_0$ but the relationships in (1.42) will remain intact, yielding $q = b_1 p_0 + d_1 i + u_1$. We thus see that b_1 , the “demand elasticity,” should be interpreted as the rate of change of Q per unit *controlled* change in P . This is different, of course, from the rate of change of Q per unit *observed* change in P (under uncontrolled conditions), which, besides b_1 , is also affected by the parameters of (1.43) (see Section 7.2.1, equation (7.14)). The difference between controlled and observed changes is essential for the correct interpretation of structural equation models in social science and economics, and it will be discussed at length in Chapter 5. If we have reasons to believe that consumer behavior will also change under a price control policy, then this modified behavior would need to be modeled explicitly—for example, by treating the coefficients b_1 and d_1 as dependent variables in auxiliary equations involving P .¹⁴ Section 7.2.1 will present an analysis of policy-related problems using this model.

To illustrate the workings of nonlinear functional models consider again the causal relationships depicted in Figure 1.2. The causal model associated with these relationships will consist of five functions, each representing an autonomous mechanism governing one variable:

$$\begin{aligned} x_1 &= u_1, \\ x_2 &= f_2(x_1, u_2), \\ x_3 &= f_3(x_1, u_3), \\ x_4 &= f_4(x_3, x_2, u_4), \\ x_5 &= f_5(x_4, u_5). \end{aligned} \tag{1.44}$$

The error variables U_1, \dots, U_5 are not shown explicitly in the graph;

¹⁴Indeed, consumers normally react to price fixing by hoarding goods in anticipation of shortages (Lucas 1976). Such phenomena are not foreign to structural models, though; they simply call for more elaborate equations to capture consumers’ expectations.

by convention, this implies that they are assumed to be mutually independent. When some disturbances are judged to be dependent, it is customary to encode such dependencies by augmenting the graph with double-headed arrows, as shown in Figure 1.1(a).

A typical specification of the functions $\{f_1, \dots, f_5\}$ and the disturbance terms is given by the following Boolean model:

$$\begin{aligned} x_2 &= [(X_1 = \text{winter}) \vee (X_1 = \text{fall}) \vee u_2] \wedge \neg u'_2, \\ x_3 &= [(X_1 = \text{summer}) \vee (X_1 = \text{spring}) \vee u_3] \wedge \neg u'_3, \\ x_4 &= (x_2 \vee x_3 \vee u_4) \wedge \neg u'_4, \\ x_5 &= (x_4 \vee u_5) \wedge \neg u'_5, \end{aligned} \tag{1.45}$$

where x_i stands for $X_i = \text{true}$ and where u_i and u'_i stand for triggering and inhibiting abnormalities, respectively. For example, u_4 stands for (unspecified) events that might cause the pavement to get wet (x_4) when the sprinkler is off ($\neg x_3$) and it does not rain ($\neg x_2$) (e.g., a broken water pipe), while u'_4 stands for (unspecified) events that would keep the pavement dry in spite of the rain (x_2), the sprinkler (x_3), and u_4 (e.g., pavement covered with a plastic sheet).

It is important to emphasize that, in the two models just described, the variables placed on the left-hand side of the equality sign (the dependent or output variables) act distinctly from the other variables in each equation. The role of this distinction becomes clear when we discuss interventions, since it is only through this distinction that we can identify which equation ought to be modified under local interventions of the type “fix the price at p_0 ” ($do(P = p_0)$) or “turn the sprinkler On” ($do(X_3 = \text{true})$).¹⁵

We now compare the features of functional models as defined in (1.40) with those of causal Bayesian networks defined in Section 1.3. Toward this end, we will consider the processing of three types of queries:

predictions (e.g., would the pavement be slippery if we *find* the sprinkler off?);

¹⁵Economists who write the supply-demand equations as $\{q = ap + u_1, q = bp + u_2\}$, with q appearing on the l.h.s. of both equations, are giving up the option of analyzing price control policies unless additional symbolic machinery is used to identify which equation will be modified by the $do(P = p_0)$ operator.

interventions (e.g., would the pavement be slippery if we *make sure* that the sprinkler is off?); and

counterfactuals (e.g., would the pavement be slippery *had* the sprinkler been off, given that the pavement is in fact not slippery and the sprinkler is on?).

We shall see that these three types of queries represent a hierarchy of three fundamentally different types of problems, demanding knowledge with increasing level of details.

1.4.2 Probabilistic Predictions in Causal Models

Given a causal model (equation (1.40)), if we draw an arrow from each member of PA_i toward X_i then the resulting graph G will be called a *causal diagram*. If the causal diagram is acyclic, then the corresponding model is called *semi-Markovian* and the values of the X variables will be uniquely determined by those of the U variables. Under such conditions, the joint distribution $P(x_1, \dots, x_n)$ is determined uniquely by the distribution $P(u)$ of the error variables. If, in addition to acyclicity, the error terms are mutually independent, the model is called *Markovian*.

A fundamental theorem about Markovian models establishes a connection between causation and probabilities via the parental Markov condition of Theorem 1.2.7.

Theorem 1.4.1 (Causal Markov Condition)

Every Markovian causal model M induces a distribution $P(x_1, \dots, x_n)$ that satisfies the parental Markov condition relative the causal diagram G associated with M ; that is, each variable X_i is independent on all its non-descendants, given its parents PA_i in G (Pearl and Verma 1991).¹⁶

The proof is immediate. Considering that the set $\{PA_i, U_i\}$ determines one unique value of X_i , the distribution $P(x_1, \dots, x_n, u_1, \dots, u_n)$ is certainly Markov relative the augmented DAG $G(X, U)$, in which the U

¹⁶Considering its generality and transparency, I would not be surprised if some version of this theorem has appeared earlier in the literature.

variables are represented explicitly. The required Markov condition of the marginal distribution $P(x_1, \dots, x_n)$ follows by d -separation in $G(X, U)$.

Theorem 1.4.1 shows that the Markov condition of Theorem 1.2.7 follows from two causal assumptions: (1) our commitment to include in the model (not in the background) every variable that is a cause of two or more other variables, and (2) Reichenbach’s (1956) common-cause assumption, also known as “no correlation without causation,” stating that, if any two variables are dependent, then one is a cause of the other *or* there is a third variable causing both. These two assumptions imply that the background factors in U are mutually independent and hence that the causal model is Markovian. Theorem 1.4.1 explains both why Markovian models are so frequently assumed in causal analysis and why the parental Markov condition (Theorem 1.2.7) is so often regarded as an inherent feature of causal models (see e.g. Kiiveri et al. 1984; Spirtes et al. 1993).¹⁷

The causal Markov condition implies that characterizing each child-parent relationship as a deterministic function, instead of the usual conditional probability $P(x_i|pa_i)$, imposes equivalent independence constraints on the resulting distribution and leads to the same recursive decomposition that characterizes Bayesian networks (see equation (1.33)). More significantly, this holds regardless of the choice of functions $\{f_i\}$ and regardless of the error distributions $P(u_i)$. Thus, we need not specify in advance the functional form of $\{f_i\}$ or the distributions $P(u_i)$; once we measure (or estimate) $P(x_i|pa_i)$, all probabilistic properties of a Markovian causal model are determined, regardless of the mechanism that actually generates those conditional probabilities. Druzdzel and Simon (1993) showed that, for every Bayesian network G characterized by a distribution P (as in (1.33)), there exists a functional model (as

¹⁷Kiiveri et al.’s (1984) paper, entitled “Recursive Causal Models,” provides the first proof (for strictly positive distributions) that the parental Markov condition of Theorem 1.2.7 follows from the factorization of (1.33). This implication, however, is purely probabilistic and invokes no aspect of causation. In order to establish a connection between causation and probability we must first devise a model for causation, either in terms of manipulations (as in Definition 1.3.1) or in terms of functional relationships in structural equations (as in Theorem 1.4.1).

in (1.40)) that generates a distribution identical to P .¹⁸ It follows that in all probabilistic applications of Bayesian networks—including statistical estimation, prediction, and diagnosis—we can use an equivalent functional model as specified in (1.40), and we can regard functional models as just another way of encoding joint distribution functions.

Nonetheless, the causal-functional specification has several advantages over the probabilistic specification, even in purely predictive (i.e. nonmanipulative) tasks. First and foremost, all the conditional independencies that are displayed by the causal diagram G are guaranteed to be *stable*—that is, invariant to parametric changes in the mechanisms represented by the functions f_i and the distributions $P(u_i)$. This means that agents who choose to organize knowledge using Markovian causal models can make reliable assertions about conditional independence relations without assessing numerical probabilities—a common ability among humanoids¹⁹ and a useful feature for inference. Second, the functional specification is often more meaningful and natural, and it yields a small number of parameters. Typical examples are the linear structural equations used in social science and economics (see Chapter 5), and the “noisy OR gate” that has become quite popular in modeling the effect of multiple dichotomous causes (Pearl 1988b, p. 184). Third (and perhaps hardest for an empiricist to accept), judgmental assumptions of conditional independence among observable quantities are simplified and made more reliable in functional models, because such assumptions are cast directly as judgments about the presence or absence of *unobserved* common causes (e.g., why is the price of beans in China judged to be independent of the traffic in Los Angeles?). In the construction of Bayesian networks, for example, instead of judging whether each variable is independent of all its nondescendants (given its parents), we need to judge whether the parent set contains *all* relevant immediate causes—in particular, whether no factor omitted from the parent set is a cause of another observed variable. Such judgments

¹⁸In Chapter 9 we will show that, except in some pathological cases, there actually exist an infinite number of functional models with such a property.

¹⁹Statisticians who are reluctant to discuss causality yet have no hesitation expressing background information in the form of conditional independence statements would probably be shocked to realize that such statements acquire their validity from none other but the *causal* Markov condition (Theorem 1.4.1). See note 9.

are more natural because they are discernible directly from a qualitative causal structure, the very structure that our mind has selected for storing stable aspects of experience.

Finally, there is an additional advantage to basing prediction models on causal mechanisms that stems from considerations of stability (Section 1.3.2). When some conditions in the environment undergo change, it is usually only a few causal mechanisms that are affected by the change; the rest remain unaltered. It is simpler then to reassess (judgmentally) or reestimate (statistically) the model parameters knowing that the corresponding symbolic change is also local, involving just a few parameters, than to reestimate the entire model from scratch.²⁰

1.4.3 Interventions and Causal Effects in Functional Models

The functional characterization $x_i = f_i(pa_i, u_i)$, like its stochastic counterpart, provides a convenient language for specifying how the resulting distribution would change in response to external interventions. This is accomplished by encoding each intervention as an alteration on a select set of functions instead of a select set of conditional probabilities. The overall effect of the intervention can then be predicted by modifying the corresponding equations in the model and using the modified model to compute a new probability function. Thus, all features of causal Bayesian networks (Section 1.3) can be emulated in Markovian functional models.

For example, to represent the action “turning the sprinkler On” in the model of (1.44), we delete the equation $x_3 = f_3(x_1, u_3)$ and replace it with $x_3 = \text{On}$. The modified model will contain all the information needed for computing the effect of the action on other variables. For example, the probability function induced by the modified model will be equal to that given by (1.36), and the modified diagram will coincide with that of Figure 1.4.

More generally, when an intervention forces a subset X of variables to attain fixed values x , then a subset of equations is to be pruned

²⁰To the best of my knowledge, this aspect of causal models has not been studied formally; it is suggested here as a research topic for students of adaptive systems.

from the model in (1.40), one for each member of X , thus defining a new distribution over the remaining variables that characterizes the effect of the intervention and coincides with the truncated factorization obtained by pruning families from a causal Bayesian network (equation (1.37)).²¹

The functional model's representation of interventions offers greater flexibility and generality than that of a stochastic model. First, the analysis of interventions can be extended to cyclic models, like the one in Figure 1.5, so as to answer policy-related questions²² (e.g.: What would the demand quantity be if we control the price at p_0 ?). Second, interventions involving the modification of equational parameters (like b_1 and d_1 in (1.42)) are more readily comprehended than those described as modifiers of conditional probabilities, perhaps because stable physical mechanisms are normally associated with equations and not with conditional probabilities. Conditional probabilities are perceived to be derivable from, not generators of, joint distributions. Third, the analysis of causal effects in non-Markovian models will be greatly simplified using functional models. The reason is: there are infinitely many conditional probabilities $P(x_i|pa_i)$ but only a finite number of functions $x_i = f_i(pa_i, u_i)$ among discrete variables X_i and PA_i . This fact will enable us in Chapter 8 (Section 8.2.2) to use linear programming techniques to obtain sharp bounds on causal effects in studies involving noncompliance.

Finally, functional models permit the analysis of context-specific actions and policies. The notion of causal effect was defined so far is of only minor use in practical policy making. The reason is that causal effects tell us the general tendency of an action to bring about a response (as with the tendency of a drug to enhance recovery in the overall population) but are not specific to actions in a given situation characterized by a set of particular observations that may themselves be affected by the action. A physician is usually concerned with the effect

²¹An explicit translation of interventions to “wiping out” equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970) and Sobel (1990). More elaborate types of interventions, involving conditional actions and stochastic strategies, will be formulated in Chapter 4.

²²Such questions, especially those involving the control of endogenous variables, are conspicuously absent from econometric textbooks (see Chapter 5).

of a treatment on a patient who has already been examined and found to have certain symptoms. Some of those symptoms will themselves be affected by the treatment. Likewise, an economist is concerned with the effect of taxation in a given economical context characterized by various economical indicators, which (again) will be affected by taxation if applied. Such context-specific causal effects cannot be computed by simulating an intervention in a static Bayesian network, because the context itself varies with the intervention and so the conditional probabilities $P(x_i|pa_i)$ are altered in the process. However, the functional relationships $x_i = f_i(pa_i, u_i)$ remain invariant, which enables us to compute context-specific causal effects as outlined in the next section (see Sections 7.2.1, 8.3, and 9.3.4 for full details).

1.4.4 Counterfactuals in Functional Models

We now turn to the most distinctive characteristic of functional models—the analysis of *counterfactuals*. Certain counterfactual sentences, as we remarked before, cannot be defined in the framework of stochastic causal networks. To see the difficulties, let us consider the simplest possible causal Bayesian network consisting of a pair of independent (hence unconnected) binary variables X and Y . Such a network ensues, for example, in a controlled (i.e. randomized) clinical trial when we find that a treatment X has no effect on the distribution of subjects' response Y , which may stand for either recovery ($Y = 0$) or death ($Y = 1$). Assume that a given subject, Joe, has taken the treatment and died; we ask whether Joe's death occurred *because of* the treatment, *despite* the treatment, or *regardless of* the treatment. In other words, we ask for the probability Q that Joe would have died had he not been treated.

To highlight the difficulty in answering such counterfactual questions, let us take an extreme case where 50% of the patients recover and 50% die in both the treatment and the control groups; assume further that the sample size approaches infinity, thus yielding

$$P(y|x) = 1/2 \quad \text{for all } x \text{ and } y. \quad (1.46)$$

Readers versed in statistical testing will recognize immediately the impossibility of answering the counterfactual question from the available

data, noting that Joe, who took the treatment and died, was never tested under the no-treatment condition. Moreover, the difficulty does not stem from addressing the question to a particular individual, Joe, for which we have only one data point. Rephrasing the question in terms of population frequencies—asking what percentage Q of subjects who died under treatment would have recovered had they not taken the treatment—will encounter the same difficulties because none of those subjects was tested under the no-treatment condition. Such difficulties have prompted some statisticians to dismiss counterfactual questions as metaphysical and to advocate the restriction of statistical analysis to only those questions that can be answered by direct tests (Dawid 1997).

However, that our scientific, legal, and ordinary languages are loaded with counterfactual utterances indicates clearly that counterfactuals are far from being metaphysical; they must have definite testable implications and must carry valuable substantive information. The analysis of counterfactuals therefore represents an opportunity to anyone who shares the aims of this book: integrating substantive knowledge with statistical data so as to refine the former and interpret the latter. Within this framework, the counterfactual issue demands answers to tough, yet manageable technical questions: What is the empirical content of counterfactual queries? What knowledge is required to answer those queries? How can this knowledge be represented mathematically? Given such representation, what mathematical machinery is needed for deriving the answers?

Chapter 7 (Section 7.2.2) presents an empirical explication of counterfactuals as claims about the temporal persistence of certain mechanisms. In our example, the response to treatment of each (surviving) patient is assumed to be persistent. If the outcome Y were a reversible condition, rather than death, then the counterfactual claim would translate directly into predictions about response to future treatments. But even in the case of death, the counterfactual quantity Q implies not merely a speculation about the hypothetical behavior of subjects who died but also a testable claim about surviving untreated subjects under subsequent treatment. We leave it as an exercise for the reader to prove that, based on (1.46) and barring sampling variations, the percentage Q of deceased subjects from the treatment group who

would have recovered had they not taken the treatment precisely equals the percentage Q' of surviving subjects in the nontreatment group who will die if given treatment.²³ Whereas Q is hypothetical, Q' is unquestionably testable.

Having sketched the empirical interpretation of counterfactuals, our next step in this introductory chapter is the question of representation: What knowledge is required to answer questions about counterfactuals? And how should this knowledge be formulated so that counterfactual queries be answered quickly and reliably? That such representation exists is evident by the swiftness and consistency with which people distinguish plausible from implausible counterfactual statements. Most people would agree that President Clinton's place in history would be different had he not met Monica Lewinsky, but only a few would assert that his place in history would change had he not eaten breakfast yesterday. In the cognitive sciences, such consistency of opinion is as close as one can get to a proof that an effective machinery for representing and manipulating counterfactuals resides someplace in the human mind. What then are the building blocks of that machinery?

A straightforward representational scheme would (i) store counterfactual knowledge in the form of counterfactual premises and (ii) derive answers to counterfactual queries using some logical rules of inference capable of taking us from premises to conclusions. This approach has indeed been taken by the philosophers Robert Stalnaker (1968) and David Lewis (1973a,b), who constructed logics of counterfactuals using closest-world semantics (i.e., " B would be true if it were A " just in case B is true in the closest possible world (to ours) in which A is true). However, the closest-world semantics still leaves two questions unanswered. (1) What choice of distance measure would make counterfactual reasoning compatible with ordinary conception of cause and effect? (2) What mental representation of interworld distances would render the computation of counterfactuals manageable and practical

²³For example, if Q equals 100% (i.e. all those who took the treatment and died would have recovered had they not taken the treatment) then all surviving subjects from the nontreatment group will die if given treatment (again, barring sampling variations). Such exercises will become routine when we develop the mathematical machinery for analyzing probabilities of causes (see Chapter 9, Theorem 9.2.11, equations (9.11)–(9.12)).

(for both humans and machines)? These two questions are answered by the structural model approach expanded in Chapter 7.

An approach similar to Lewis's (though somewhat less formal) has been pursued by statisticians in the potential-outcome framework (Rubin 1974; Robins 1986; Holland 1988). Here, substantive knowledge is expressed in terms of probabilistic relationships (e.g. independence) among counterfactual variables and then used in the estimation of causal effects. The question of representation shifts from the closest-world to the potential-outcome approach: How are probabilistic relationships among counterfactuals stored or inferred in the investigator's mind? In Chapter 7 (see also Section 3.6.3) we provide an analysis of the closest-world and potential-outcome approaches and compare them to the structural model approach, to be outlined next, in which counterfactuals are *derived* from (and in fact defined by) a functional causal model (equation (1.40)).

In order to see the connection between counterfactuals and structural equations, we should first examine why the information encoded in a Bayesian network, even in its causal interpretation, is insufficient to answer counterfactual queries. Consider again our example of the

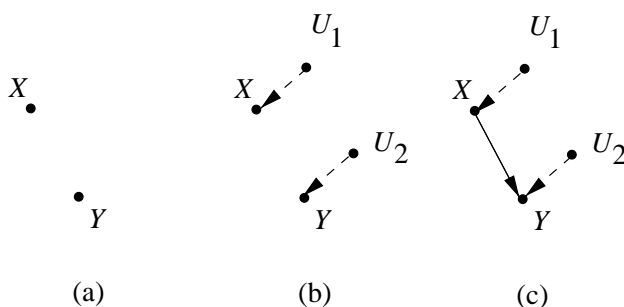


Figure 1.6: (a) A causal Bayesian network that represents the distribution of (1.47). (b) A causal diagram representing the process generating the distribution in (a), according to model 1. (c) Same, according to model 2. (Both U_1 and U_2 are unobserved.)

controlled randomized experiment (equation 1.46), which corresponds to an edgeless Bayesian network (Figure 1.6(a)) with two independent

binary variables and a joint probability:

$$P(y, x) = 0.25 \quad \text{for all } x \text{ and } y. \quad (1.47)$$

We now present two functional models, each generating the joint probability of (1.47) yet each giving a different value to the quantity of interest, Q = the probability that a subject who died under treatment ($x = 1$, $y = 1$) would have recovered ($y = 0$) had he or she not been treated ($x = 0$).

Model 1 (Figure 1.6(b))

Let

$$\begin{aligned} x &= u_1, \\ y &= u_2, \end{aligned}$$

where U_1 and U_2 are two independent binary variables with $P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2}$ (e.g., random coins).

Model 2 (Figure 1.6(c))

Let

$$\begin{aligned} x &= u_1, \\ y &= xu_2 + (1 - x)(1 - u_2), \end{aligned} \quad (1.48)$$

where, as before, U_1 and U_2 are two independent binary variables.

Model 1 corresponds to treatment (X) that has no effect on any of the subjects; in model 2, every subject is affected by treatment. The reason that the two models yield the same distribution is that model 2 describes a mixture of two subpopulations. In one ($u_2 = 1$), each subject dies ($y = 1$) if and only if treated; in the other ($u_2 = 0$), each subject recovers ($y = 0$) if and only if treated. The distributions $P(x, y, u_2)$ and $P(x, y)$ corresponding to these two models are shown in the tables of Figure 1.7.

The value of Q differs in these two models. In model 1, Q evaluates to zero, because subjects who died correspond to $u_2 = 1$ and, since the treatment has no effect on y , changing X from 1 to 0 would still yield $y = 1$. In model 2, however, Q evaluates to unity, because subjects who died under treatment must correspond to $u_2 = 1$ (i.e., those who die if treated), meaning they would recover if and only if not treated.

<u>Model 1</u>	$u_2 = 0$		$u_2 = 1$		Marginal	
	$x = 1$	$x = 0$	$x = 1$	$x = 0$	$x = 1$	$x = 0$
$y = 1$ (death)	0	0	0.25	0.25	0.25	0.25
$y = 0$ (recovery)	0.25	0.25	0	0	0.25	0.25

<u>Model 2</u>	$u_2 = 0$		$u_2 = 1$		Marginal	
	$x = 1$	$x = 0$	$x = 1$	$x = 0$	$x = 1$	$x = 0$
$y = 1$ (death)	0	0.25	0.25	0	0.25	0.25
$y = 0$ (recovery)	0.25	0	0	0.25	0.25	0.25

Figure 1.7: Contingency tables showing the distributions $P(x, y, u_2)$ and $P(x, y)$ for the two models discussed in the text.

The first lesson of this example is that stochastic causal models are insufficient for computing probabilities of counterfactuals; knowledge of the actual process behind $P(y|x)$ is needed for the computation.²⁴ A second lesson is that a functional causal model constitutes a mathematical object sufficient for the computation (and definition) of such probabilities. Consider, for example, model 2 of (1.48). The way we concluded that a deceased treated subject ($y = 1, x = 1$) would have recovered if not treated involved three mental steps. First, we applied the evidence at hand, $e : \{y = 1, x = 1\}$, to the model and concluded that e is compatible with only one realization of U_1 and U_2 —namely, $\{u_1 = 1, u_2 = 1\}$. Second, to simulate the hypothetical condition “had he or she not been treated,” we substituted $x = 0$ into (1.48) while ignoring the first equation $x = u_1$. Finally, we solved (1.48) for y (assuming $x = 0$ and $u_2 = 1$) and obtained $y = 0$, from which we concluded that the probability of recovery ($y = 0$) is unity under the hypothetical condition considered.

²⁴In the potential-outcome framework (Sections 3.6.3 and 7.4.4), such knowledge obtains stochastic appearance by defining distributions over *counterfactual variables* Y_1 and Y_0 , which stand for the potential response of an individual to treatment and no treatment, respectively. These hypothetical variables play a role similar to the functions $f_i(pa_i, u_i)$ in our model; they represent the deterministic assumption that every individual possesses a definite response to treatment, regardless of whether that treatment was realized.

These three steps can be generalized to any causal model M as follows. Given evidence e , to compute the probability of $Y = y$ under the hypothetical condition $X = x$ (where X is a subset of variables), apply the following three steps to M .

Step 1 (abduction): Update the probability $P(u)$ to obtain $P(u|e)$.

Step 2 (action): Replace the equations corresponding to variables in set X by the equations $X = x$.

Step 3 (prediction): Use the modified model to compute the probability of $Y = y$.

In temporal metaphors, this three-step procedure can be interpreted as follows. Step 1 explains the past (U) in light of the current evidence e ; step 2 bends the course of history (minimally) to comply with the hypothetical condition $X = x$; finally, step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

Recalling that for each value u of U there is a unique solution for Y , it is clear that step 3 always gives a unique solution for the needed probability; we simply sum up the probabilities $P(u|e)$ assigned to all those u that yield $Y = y$ as a solution. Chapter 7 develops effective procedures for computing probabilities of counterfactuals, procedures that are based on probability propagation in “twin” networks (Balke and Pearl 1995): one network represents the actual world; the other the counterfactual world.

Note that the hypothetical condition $X = x$ always stands in contradiction to the prevailing values u of U in the model considered (else $X = x$ would actually be realized and thus would not be considered hypothetical). It is for this reason that we invoke (in step 2) an external intervention (alternatively, a “theory change” or a “miracle”; Lewis 1973b), which modifies the model and thus explains the contradiction away. In Chapter 7 we extend this structural-interventional model to give a full semantical and axiomatic account both for counterfactuals and the probability of counterfactuals. In contrast with Lewis’s theory, this account is not based on abstract notion of similarity among hypothetical worlds; rather, it rests on the actual mechanisms involved

in the production of the hypothetical worlds considered. Likewise, in contrast with the potential-outcome framework, counterfactuals in the structural account are not treated as undefined primitives but rather as quantities to be derived from the more fundamental concepts of causal mechanisms and their structure.

The three-step model of counterfactual reasoning also uncovers the real reason why stochastic causal models are insufficient for computing probabilities of counterfactuals. Because the U variables do not appear explicitly in stochastic models, we cannot apply step 1 so as to update $P(u)$ with the evidence e at hand. This implies that several ubiquitous notions based on counterfactuals—including probabilities of causes (given the effects), probabilities of explanations, and context-dependent causal effect—cannot be defined in such models. For these, we must make some assumptions about the form of the functions f_i and the probabilities of the error terms. For example, the assumptions of linearity, normality, and error independence are sufficient for computing all counterfactual queries in the model of Figure 1.5 (see Section 7.2.1). In Chapter 9, we will present conditions under which counterfactual queries concerning probability of causation can be inferred from data when f_i and $P(u)$ are unknown, and only general features (e.g. monotonicity) of these entities are assumed. Likewise, Chapter 8 (Section 8.3) will present methods of *bounding* probabilities of counterfactuals when only stochastic models are available.

The preceding considerations further imply that the three tasks listed in the beginning of this section—prediction, intervention, and counterfactuals—form a natural hierarchy of causal reasoning tasks, with increasing levels of refinement and increasing demands on the knowledge required for accomplishing these tasks. Prediction is the simplest of the three, requiring only a specification of a joint distribution function. The analysis of interventions requires a causal structure in addition to a joint distribution. Finally, processing counterfactuals is the hardest task because it requires some information about the functional relationships and/or the distribution of the omitted factors.

This hierarchy also defines a natural partitioning of the chapters in this book. Chapter 2 will deal primarily with the probabilistic aspects of causal Bayesian networks (though the underlying causal structure will serve as a conceptual guide). Chapters 3–6 will deal exclusively

with the interventional aspects of causal models, including the identification of causal effects, the clarification of structural equation models, and the relationships between confounding and collapsibility. Chapters 7–10 will deal with counterfactual analysis, including axiomatic foundation, applications to policy analysis, the bounding of counterfactual queries, the identification of probabilities of causes, and the explication of single-event causation.

I wish the reader a smooth and rewarding journey through these chapters. But first, an important stop for terminological distinctions.