2.7 Local Criteria for 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., $X \perp\!\!\!\perp Z|S$) and*
   (ii) *Z and Y are dependent given S (i.e., $Z \not\perp\!\!\!\perp Y|S$).*

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 genuine cause of d

(with $Z = b$ and $S = a$). Neither b nor c qualifies as 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., $Z \not\perp\!\!\!\perp Y | S$), and*

   (iii) *Z and Y are independent given $S \cup X$ (i.e., $Z \perp\!\!\!\perp Y | S \cup X$);*

   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 $P(Y|X) > P(Y)$ (in the spirit of Reichenbach 1956; Suppes 1970; and Good 1961). This refinement is applicable to any of the definitions in this section, but it will not be formulated explicitly.

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

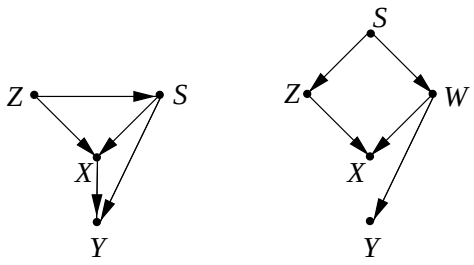
Conditions 1 and 2 use Z_1 and S_1 to disqualify Y as a cause of X, paralleling conditions (i)–(ii) of Definition 2.7.1; conditions 3 and 4 use Z_2 and S_2 to disqualify X as a cause of Y. This leaves the existence of a latent common cause as the only explanation for the observed dependence between X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

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 cause 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. *$(Z \not\perp\!\!\!\perp Y|S)$;*
2. *$(Z \perp\!\!\!\perp Y|S \cup X)$.*

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.



(a) (b)

Figure 2.5: Illustrating 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.

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

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 $(a \perp\!\!\!\perp b | \emptyset)$, $(a \not\perp\!\!\!\perp c | \emptyset)$, and $(b \not\perp\!\!\!\perp c | \emptyset)$. The argument goes as follows. If we

[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.

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 $a \leftrightarrow c \leftrightarrow b$. 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 triples 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 1998a, 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).