Implementation-Neutral Causation

Stephen F. LeRoy
University of California, Santa Barbara

October 24, 2013

Abstract

The most basic question one can ask of a model is “According to the model, what is the effect of variable \(y_1\) on variable \(y_2\)?”, where \(y_1\) and \(y_2\) are two variables of the model. Causation is “implementation neutral” when all interventions on external variables that lead to a given change in \(y_1\) have the same effect on \(y_2\). We derive conditions for implementation neutrality in simple formal models. From there the discussion turns to empirical aspects of implementation neutrality. As an application of the analysis, the conditions for implementation-neutral causation are compared with those required for Granger non-causation.

The most basic question one can ask of a model is “According to the model, what is the effect of variable \(y_1\) on variable \(y_2\)?”, where \(y_1\) and \(y_2\) are two variables of the model.\(^1\) Two answers are possible. The first involves observing that many possible interventions on the model’s external variables could have led to the assumed change in \(y_1\), and in general the effects of these interventions on \(y_2\) are different. Therefore the question “What is the effect of \(y_1\) on \(y_2\)?” does not have an unambiguous answer: the information given about the intervention—its effect on \(y_1\)—is insufficient to characterize the effect of the intervention on \(y_2\).

The second answer is that even though the intervention is not completely characterized for the reason just noted, all interventions consistent with the

\(^1\)Abbreviated versions of this material were presented in LeRoy [11] and [12] (these papers are discussed at some length in Cartwright [1]). Here more detail is supplied, and the packaging is somewhat different.
assumed change in $y_1$ may map onto the same change in $y_2$. In that case the question "What is the effect of $y_1$ on $y_2$?" has a well-defined answer. In linear systems, to which our attention will be restricted in this paper, the effect is captured by a single constant, here labeled $a_{21}$, which gives the effect on $y_2$ of a unit change of $y_1$, regardless of what intervention caused the change in $y_1$.

If, as in the second case above, the effect of a change in $y_1$ on $y_2$ is independent of how the change in $y_1$ is implemented, we will say that the intervention is implementation-neutral, and will write $y_1 \rightarrow y_2$. Thus, as the notation indicates, we incorporate implementation neutrality in the definition of causation. We do this on the grounds that implementation neutrality is implicit in the idea of causation: if implementation neutrality fails one cannot evaluate quantitatively the effect of $y_1$ on $y_2$, that not being implied by the assumed change in $y_1$. If implementation neutrality fails one can only analyze the effect on $y_2$ of various specific interventions on the determinants of $y_1$. Doing so amounts to abandoning the attempt to characterize $y_1$ as a cause of $y_2$, instead shifting the discussion to treating the determinants of $y_1$ as the causes of $y_2$. Restricting the meaning of causation to cases where interventions are implementation neutral is, of course, purely a matter of definition: if one prefers to say instead that $y_1$ causes $y_2$ but only an implementation-specific account of the magnitude of the effect is available, no harm is done, although doing so seems pointless to the extent that one is interested in evaluating causation quantitatively.

The question of implementation neutrality is of central importance in applied scientific work: analysts want either to measure the impact of the treatment on the patient or to know that doing so is impossible because the effect of that treatment depends on how it is applied. Restricting the term "causation" to settings where causation is implementation neutral amounts only to rejecting application of the term in settings in which it does not have a clear meaning.\footnote{Note that throughout this paper no notational distinction is made between the name of a variable and the values that variable takes on.} \footnote{However, the opposite can be argued. Pearl [13], p. 136, took the view that linking causation with implementation neutrality "denies any causal reading to most of the structural parameters that economists and social scientists labor to estimate." On the contrary, if the argument here is accepted attempts by economists and social scientists to estimate parameters associated with causation are hopeless when implementation neutrality fails because in that case these parameters are not well defined.} Granted, the proposed reading of causation is not
the only possible one, and it would not be appropriate to insist on rejecting application of the term when implementation neutrality fails. That said, it is far from clear that there is anything worthwhile to be said about causation if implementation neutrality fails, as noted in the preceding paragraph.

Use of graphical methods in causal analysis has become widespread in recent years, due to work by Pearl [13], Spirtes, Glymour and Scheines [14], Woodward [16], Hausman [7], Cartwright [1] and others. These authors typically do not include implementation neutrality in their definition of causation. As we will see, implementation-neutral causation is antisymmetric, so it results in directed acyclic graphs of the type in common use. Therefore one has the option of imposing implementation neutrality in the derivation of causal orderings represented as directed acyclic graphs, and comparing how doing so affects the results. This paper addresses questions related to the construction and interpretation of directed acyclic graphs when causation is restricted to be implementation neutral.

1 Characterization of Implementation-Neutral Causation

A distinction that is central in any model that deals with issues of causation is that between internal and external variables. Internal variables are those determined by the model, while external variables are those taken as given; i.e., determined outside the model.\footnote{In the earlier literature the terms “endogenous” and “exogenous” were often used in place of “internal” and “external”. The earlier usage is consistent with the etymology of the terms, but econometricians have implemented a change in their meaning (see Granger [6], wherein I am reprimanded for using the earlier terms in LeRoy [11]). To avoid ambiguity, economists now use “internal” and “external” when the earlier meaning is intended, as here.

For discussion of various definitions of exogeneity and endogeneity see Leamer [10]. For a statement of the definition of exogeneity currently favored by econometricians see Engle, Hendry and Richard [4].} We will use $y$ to denote internal variables and $x$ to denote external variables. All changes in solution values of internal variables are assumed to be attributable to interventions on external variables, as opposed to alterations of equations. Doing so requires the analyst to be explicit about which hypothetical alterations in the model are permitted and which are ruled out. External variables are assumed to
be variation-free: that is, the analyst is free to alter them independently. Independent variation corresponds to the assumption that by definition external variables are not linked by functional relations; otherwise they would be classified as internal. The solution form of a model expresses each internal variable as a function of the set of external variables that determine it.\(^5\)

We will refer to the set of external variables that determine any internal variable as its \textit{external set}, and will denote the external set for \(y_i\) as \(\mathcal{E}(y_i)\). We will assume that the external set for any internal variable consists of at least two external variables. Otherwise the internal variable is a rescaling of the external variable (assuming linearity); a model containing an internal variable the external set of which consists of one external variable can be simplified by deleting one of the variables.

There is no ambiguity in defining causation when the cause variable is external: \(x_i\) causes \(y_j\) whenever \(x_i\) is in the external set for \(y_j\) and, by virtue of linearity, in that case a unique constant \(b_{ji}\) gives the effect of a unit change in \(x_i\) on \(y_j\) for any values of the external variables. If \(x_i\) is not in the internal set for \(y_j\) the former does not cause the latter. The ambiguity comes up when the cause variable is internal, since then an assumed change in the cause variable could come from interventions in any or all of the variables in its external set, and in general the effect of the interventions on \(y_j\) is different for each possible set of interventions. This is so even if all the contemplated interventions are restricted to have the same effect on the cause variable. Given this ambiguity, we cannot attribute causation in this case: the intervention is not described with sufficient detail to generate a clear answer.

However, consider a special case in which two conditions on the external sets of two variables and the functions mapping these sets to the two variables are satisfied. The first is the \textit{subset condition}: the external set for \(y_i\) is a proper subset of that of \(y_j\). The subset condition guarantees that any external variable that affects \(y_i\) also affects \(y_j\), but not vice-versa. Hoover [9] in particular emphasized this condition, which assures the antisymmetry of causation.\(^6\)

\(^5\)We thus distinguish between the solution form and the reduced form, in which current-date internal variables are expressed as functions of lagged internal variables and external variables. In static models the solution form and reduced form coincide.

\(^6\)In this paper the subset condition is a condition we impose on models to assure that causation is antisymmetric. Hausman ([7], Ch. 4) had a different take on what we call the subset condition. Hausman’s \textit{independence condition} states that “if \(a\) causes \(b\) ..., then \(b\) has a cause that is distinct from \(a\) and not causally connected to \(a\).” Hausman appears to
The second is the sufficiency condition, which states that the map from $E(y_j)$ to $y_j$ can be expressed as the sum of two functions. The first function is the composition of a function from $E(y_i)$ to $y_i$ and a function from $y_i$ to $y_j$, while the second is a function from $E(y_j) - E(y_i)$ to $y_j$. If such functions exist, $y_i$ is a sufficient statistic for $E(y_i)$ for the purpose of determining $y_j$, meaning that for the purpose of determining $y_j$ an intervention on any or all of the variables in $E(y_i)$ is adequately characterized by the resulting induced change in $y_i$. If the subset and sufficiency conditions are satisfied we can write $y_i \to y_j$. In that case causation is implementation neutral, since any intervention on the elements of $E(y_i)$ that results in a fixed $\Delta y_i$ will result in the same $\Delta y_j$.

Heckman [8], among others, pointed out that the received graphical treatment of causation applies only to recursive models (the graphical treatment outlined here, in contrast, can be applied to nonrecursive models). Heckman rejected the view that “the econometric analysis” of causation fails in nonrecursive causal models. However, it turns out that he was associating analysis of causation with recovery of structural parameters, which does not involve a precise characterization of causation. For example, he correctly observed that in the supply-demand model the price elasticities of supply and demand are identified under the appropriate exclusion restrictions. He also noted, again correctly, that there is no ambiguity in characterizing the causal effects of the external variables on internal variables. These facts, however, do not alter the point that price and quantity, being simultaneously determined, are themselves not causally ordered.

In the causal form of a model the equations are written in such a way as to incorporate its causal structure. Starting from the solution form of the model, one can readily derive its causal form. First one derives the causal ordering, which consists of determining for each $i$ and $j$ whether or not we have $y_i \to y_j$. In the causal form of the model the equation for each internal

---

view the independence condition, not as an assumption in a model, but as a proposition about the world that may or may not be true: “As a metaphysical claim about patterns of lawlike connections found in nature, [the independence condition] seems incredible, and its truth miraculous.” (p. 64).

However, he went on to consider another possible interpretation, that the failure of the independence condition implies only that there may exist lawlike relations in the world that are not specifically causal relations. This is so because causality inherently involves antisymmetry, and antisymmetry may not occur if the independence condition fails. This latter interpretation, which relates the independence condition to the definition of independence, is closer to the position taken here.
variable $y_i$ that has no internal variables as causal parents coincides with the corresponding equation in the solution form of the model (i.e., consists of a map from $\mathcal{E}(y_i)$ to $y_i$). The causal form for internal variables $y_j$ that have one or more causal parents consists of a map from the parent, or from each of the parents, to $y_j$, plus a map to $y_j$ from the elements of $\mathcal{E}(y_j)$ that are not in the external sets of any of the parents of $y_j$:

$$y_j \leftarrow f_j(y_i, x_k).$$

Here $y_i$ is the (single, in this case) internal variable that causes $y_j$, and $x_k = \mathcal{E}(y_j) - \mathcal{E}(y_i)$. The case in which $y_j$ has more than one cause is handled by expanding the set of arguments of $f_j$ appropriately. Note our substitution of $\leftarrow$ for $=$; since causation is irreflexive and antisymmetric it is inappropriate to use the equality relation in writing the causal form of a model.

Observe that the conditions for causation differ from those defining recursive models. Any model can be put in recursive form by using algebraic operations on the model’s equations, but the causal form depends on assumed parameter restrictions. It follows that writing a model in recursive form does not imply that the coefficients can be interpreted as quantifying a causal ordering.

One can represent the causal form of the model by a causal graph. For variables $y_i$ without parents this consists of arrows drawn to $y_i$ from each element of $\mathcal{E}(y_i)$. For variables with parents the arrows run to $y_j$ from the parent(s) of $y_j$, and also to $y_j$ from each variable in the external set of $y_j$ but not in the external sets of any of its parents.

An example will make this clear.

1.1 Example

Consider the following model, written in solution form:

$$y_1 = b_{11}x_1 + b_{12}x_2$$  \hspace{1cm} (1)

$$y_2 = b_{21}x_1 + b_{22}x_2 + b_{23}x_3.$$  \hspace{1cm} (2)

The external sets for $y_1$ and $y_2$ are $\mathcal{E}(y_1) = \{x_1, x_2\}$ and $\mathcal{E}(y_2) = \{x_1, x_2, x_3\}$. The former is a strict subset of the latter, so the subset condition for $y_1 \to y_2$ is satisfied. Without parameter
restrictions the sufficiency condition for \( y_1 \to y_2 \) is not satisfied. However, if the condition
\[
b_{21}/b_{11} = b_{22}/b_{12}
\] (3)
obtains the sufficiency condition is satisfied: in that case we can define \( a_{21} \) by
\[
a_{21} \equiv b_{21}/b_{11} = b_{22}/b_{12},
\] (4)allowing replacement of (2) by
\[
y_2 = a_{21}y_1 + b_{23}x_3
\] (5)(use (3) and (1) to eliminate \( a_{21} \) and \( y_1 \) in (5), resulting in (2)). We have \( y_1 \to y_2 \). To repeat, \( y_1 \) causes \( y_2 \) in this case because all interventions in \( x_1 \) and \( x_2 \) consistent with a given change in \( y_1 \) have the same effect on \( y_2 \). If, on the other hand, the condition (3) fails then different interventions in \( x_1 \) and \( x_2 \) consistent with a given change in \( y_1 \) lead to different effects on \( y_2 \), so the outcome on \( y_2 \) is ambiguous. In that case we will say that \( y_1 \) does not cause \( y_2 \).

If the condition (3) is satisfied the causal form of the model consists of equations (1) and (5), while if it is not the causal form coincides with the solution form, which is (1) and (2).

The left-hand panel of Figure 1 shows the causal graph if the restriction (3) is satisfied; the right-hand panel shows the causal graph if the restriction is not satisfied. □

The subset condition guarantees that if \( y_i \to y_j \) then there exists at least one external variable that causes \( y_j \) but does not cause \( y_i \). This condition
implies that causation is antisymmetric \((y_i \rightarrow y_j \text{ implies } y_j \not\leftrightarrow y_i)\). This does not mean that it is necessary to restrict attention to models that are recursive; however, it is the case that to the extent that models contain blocks of simultaneous equations there will not exist causal arrows connecting the variables in these blocks. In the extreme case of a completely simultaneous model without parameter restrictions the causal ordering will be empty and the causal form of the model will be the same as the solution form. This makes sense: in models without special restrictions the effect of one internal variable on another is ambiguous.

An important aspect of the definition of causation just presented is that causation is transitive: if \(y_1 \rightarrow y_2\) with coefficient \(a_{21}\) and \(y_2 \rightarrow y_3\) with coefficient \(a_{32}\), then \(y_1 \rightarrow y_3\) with coefficient \(a_{31} = a_{21}a_{32}\). The transitivity of causation has the implication that an internal variable never has both an indirect effect on another variable via a causal chain involving one or more third variables, and also a distinct direct effect; rather, the direct effect is always the composition of the indirect effects. This outcome, although somewhat counterintuitive, is an implication of implementation neutrality. Observe that the converse of this transitivity property does not obtain: if we maintain \(y_1 \rightarrow y_2\), then \(y_1 \rightarrow y_3\) does not imply \(y_2 \rightarrow y_3\). This conclusion agrees with common sense—the fact that \(y_2\) and \(y_3\) have \(y_1\) as a common parent does not imply that \(y_2\) is a parent of \(y_3\).

The transitivity of causation has an implication for causal graphs. Suppose we have \(y_1 \rightarrow y_2 \rightarrow y_3\). Then the causal path that connects \(y_1\) and \(y_3\) consists of a path from \(y_1\) to \(y_2\) linked to a path from \(y_2\) to \(y_3\). Conversely, if there exists a path from \(y_1\) to \(y_3\) that does not go through \(y_2\), we cannot have \(y_1 \rightarrow y_2 \rightarrow y_3\).

An expanded version of the example will illustrate these relations.

1.2 Example

Consider the model

\[
y_1 = b_{11}x_1 + b_{12}x_2 \tag{6}
\]
\[
y_2 = b_{21}x_1 + b_{22}x_2 + b_{23}x_3 \tag{7}
\]
\[
y_3 = b_{31}x_1 + b_{32}x_2 + b_{33}x_3 + b_{34}x_4. \tag{8}
\]
Figure 2:

Figure 3:
We have already noted that if we have $b_{21}/b_{11} = b_{22}/b_{12}$ we can define $a_{21}$ by $a_{21} \equiv b_{21}/b_{11} = b_{22}/b_{12}$ and can replace (7) by

$$y_2 = a_{21}y_1 + b_{23}x_3.$$  \(\text{(9)}\)

In that case we have $y_1 \rightarrow y_2$ (Figure 2). Similarly, if we have $b_{31}/b_{11} = b_{32}/b_{12}$ we can replace (8) by

$$y_3 = a_{31}y_1 + b_{33}x_3 + b_{34}x_4,$$  \(\text{(10)}\)

and we have $y_1 \rightarrow y_3$ (Figure 3). If the condition $b_{31}/b_{21} = b_{32}/b_{22} = b_{33}/b_{23}$ is satisfied we have

$$y_3 = a_{32}y_2 + b_{34}x_4,$$  \(\text{(11)}\)

implying $y_2 \rightarrow y_3$ (Figure 4). If both conditions are satisfied we have that $y_1 \rightarrow y_2$ and $y_2 \rightarrow y_3$, so we also have $y_1 \rightarrow y_3$, by the transitivity of causation (Figure 5). The interpretation is that, as noted above, direct and indirect causation do not coexist as distinct causal effects: if we have both $y_1 \rightarrow y_2$ and $y_2 \rightarrow y_3$, then $y_1 \rightarrow y_3$ obtains, and it is just the composition of $y_1 \rightarrow y_2$ and $y_2 \rightarrow y_3$. Accordingly, the coefficient associated with $y_1 \rightarrow y_3$ is the product of the coefficients associated with $y_1 \rightarrow y_2$ and $y_2 \rightarrow y_3$. □

As noted above, the finding that causation from $y_1$ to $y_3$ is necessarily the composition of causation from $y_1$ to $y_2$ and $y_2$ to $y_3$ (when these occur) is counterintuitive. This is especially so for those accustomed to the received treatment of causation, in which direct causation can have a different interpretation from the composition of the direct causal links. However, again, it
is an inevitable consequence of the implication of implementation neutrality that causation is transitive.

We noted above that one cannot write down an arbitrary structural equation and interpret the coefficients of its internal explanatory variables as necessarily reflecting causation. In Figure 6 the graph indicates that we have \( y_1 \rightarrow y_2 \), but by inspection the external sets of \( y_1 \) and \( y_2 \) are the same, contradicting \( y_1 \rightarrow y_2 \). There is no problem with the model; the assertion here is only that the model coefficients cannot be associated with causation under its present definition, so the graph is invalid.

Figure 7 gives another example of a causal diagram that could not occur under the definition of causation proposed here.

To make a related point, consider again Figure 5, where arrows connect \( y_1 \) and \( y_3 \) and also \( y_2 \) and \( y_3 \). Under the received characterization of causation the graph would be interpreted as saying that the true causal relation between \( y_2 \) and \( y_3 \) is represented by the coefficient \( a_{32} \); this despite the fact that two of the three external variables that determine \( y_3 \) also determine \( y_1 \), which is
connected to $y_3$ via $y_2$. In the received treatment $x_1$ and $x_2$ are held to affect $y_3$ only insofar as they determine the value of $y_2$—in the accepted language, $y_2$ is a “confounding variable”—whereas $x_3$ is a direct cause of $y_3$.

It is difficult to see any reason for assigning a privileged position to $x_3$ here: $x_1$, $x_2$ and $x_3$ are all are causes of $y_3$ and there is no obvious motivation for relegating the first two to the class of confounding variables.

Expanding on this point, there is nothing about directed acyclic graphs that points to the distinction in the treatment of $x_3$ relative to $x_1$ and $x_2$ in the received analysis of causation: the graphs produced under the present characterization of causation are directed and acyclic, but they do not provide any justification for the confounding story. So where does the confounding story come from? In the absence of a clear rationale for the confounding story based on the received definition of causation there does not seem to be a clear way around this problem.

One always has the option of viewing $a_{32}$ as capturing the effect of $y_2$ on $y_3$ with other causes (specifically in the above example, $y_1$) held constant. We do not do this, and it is worthwhile explaining why. Most simply, this “ceteris paribus” interpretation involves fixing $x_1$ and $x_2$ and interpreting the assumed intervention as involving $x_3$ only. This is acceptable—and may be what proponents of the confounding interpretation have in mind. But in that case the relevant causation is from $x_3$ to $y_3$, not $y_2$ to $y_3$, and the coefficient involved with causation from $x_3$ to $y_3$ is $b_{33}$ ($= b_{32} a_{32}$). It is unclear why one would want to connect $y_2$ with an intervention that is completely characterized as an assumed shift in $x_3$; we are talking about the effect of $x_3$ on $y_3$, so why involve $y_2$?

A related option is to interpret the intervention as involving nonzero changes in $x_1$ and $x_2$ as well as an arbitrary change in $x_3$, but restricting the interventions in $x_1$ and $x_2$ to imply $\Delta y_2 = 0$. Again, this allows the analyst
to identify causation with \( a_{32} \). The problem with this is that postulating a functional relation between \( x_1 \) and \( x_2 \) conflicts with the assumed status of these variables as variation-free external variables. If one wishes to consider this situation it is appropriate to specify the form of the functional relation assumed to link \( x_1 \) and \( x_2 \)—i.e., specify whether \( x_1 \) causes \( x_2 \) or vice versa, for example—and alter the model by including the new equation in the formal structure.

2 Empirical Aspects of Causation

Up to this point we have considered models in which variables are specified as to their status as internal or external. We have not specified which variables are observable or what we are assuming about the probability distributions of those that are not. That we could postpone discussion of observability to this point reflects the facts that the characterization of causation and the existence or nonexistence of the constants associated with causation depend only on whether the conditions for implementation neutrality are satisfied. They do not depend on which variables are observable or what is assumed about those that are not. However, without specifying which variables are observable and characterizing the probability distribution of unobserved external variables there is no way to estimate causal coefficients: in the absence of identifying information the correlations among internal variables implied by the model’s causal structure cannot be disentangled those induced by correlations among the external variables.

The most direct way to launch an investigation of the empirical implications of causation is to assume, first, that external variables are unobservable and internal variables are observable. This restriction is not without loss of generality. Assuming that external variables are not observable rules out at least some kinds of nonstationary models. Further, assuming that internal variables are observable rules out latent variables, a topic of major interest in applied work. The motivation for adopting these restrictions is to simplify the analysis.

Second, it is assumed that the external variables are statistically independent random variables. This assumption implies that whatever correlations exist among the model’s internal variables are generated by the equations of the model, not from uninterpreted correlations among external variables.

The assumptions just listed imply that if we have \( y_1 \rightarrow y_2 \) the coefficient
measuring the effect of $y_1$ on $y_2$ is identified, and can be estimated consistently and without bias using a least-squares regression of $y_2$ on $y_1$. This is so because the external variable(s) in $E(y_2) - E(y_1)$—the constituents of the error term in the regression—is (are) independent of $E(y_1)$, and therefore of $y_1$ itself.\footnote{The converse of this result, of course, does not hold: the fact that one can always run a regression of one variable on another does not mean that the resulting coefficient can be interpreted as an estimate of a causal coefficient.}

The finding that causal coefficients are always identified differs from the conclusion of the Cowles economists. The reason for the difference is that the Cowles economists defined causation from what they called structural equations, the coefficients of which may or may not be identified. Here, in contrast, causation involves regression coefficients among internal variables. By assumption internal variables are observable, implying that causal coefficients are identified. There is a caveat: the finding that causal coefficients are always identified should not be interpreted as implying that causation is easily testable: recall that the coefficient restrictions needed for causation involve external variables, which by definition are not observable.

Under causation as characterized here, as with other definitions of causation, a model that includes an account of causation can in principle be tested by identifying pairs of variables that are or are not statistically independent according to the model, and then determining whether these independence implications are satisfied empirically.

We now consider whether powerful empirical tests of causal models along these lines are likely to be available. It appears that they are not: only in special cases is it possible to characterize independence or the lack thereof among internal variables as testable implications of causal models. Among the few results that are available is the obvious fact that any two internal variables for which the external sets are disjoint are statistically independent.

As an implication, if an internal variable has two parents, then either one parent causes the other or the two are statistically independent. To see this, suppose that $y_1 \rightarrow y_3$ and $y_2 \rightarrow y_3$, so that $y_3$ has parents $y_1$ and $y_2$. If $E(y_1)$ and $E(y_2)$ are disjoint, then $y_1$ and $y_2$ are statistically independent. Suppose instead that $E(y_1)$ and $E(y_2)$ have a nonempty intersection that contains external variable $x$. Then since $x \in E(y_1)$ there exists a path from $x$ to $y_3$ that includes $y_1$. Because $x \in E(y_2)$ there exists a path from $x$ to $y_3$ that includes $y_2$. These must be the same path, since if there were a path...
that included $y_1$ but not $y_2$ then $y_2$ could not be a sufficient statistic for $\mathcal{E}(y_2)$, contradicting $y_2 \rightarrow y_3$. Thus there is a single path connecting $x$ and $y_3$, and that path includes both $y_1$ and $y_2$. This can occur only if $y_1 \rightarrow y_2$ or $y_2 \rightarrow y_1$. This result is due directly to our association of causation with implementation neutrality.

Past this there is not much to be said. Assume that $y_1$ and $y_2$ have $y_3$ as a common parent. If also $y_1 \rightarrow y_2$, then we have $y_3 \rightarrow y_1 \rightarrow y_2$. In that case we have that all three variables are correlated since their external sets have a nonempty intersection (consisting of the external set for $y_3$). If, on the other hand, $y_1 \not\rightarrow y_2$ the causal coefficient associated with $y_1 \rightarrow y_2$ is not defined, so the question of bias in estimating that coefficient does not arise.

A related point is that even under the assumption that external variables are independently distributed, causation as characterized here does not satisfy the causal Markov condition (which says that any variable is statistically independent of any other variable that is not among the causal descendants of the first variable, conditional on the first variable’s parents). This is obviously true, for example, in fully simultaneous systems. In that case the causal ordering is empty, so the causal Markov condition implies that all internal variables are statistically independent. This is generally not the case.

Implementation neutrality implies that the causal Markov condition may fail even in systems that are not fully simultaneous. For instance, in the example of Subsection 1.2 with $y_1 \rightarrow y_2$ but not $y_2 \rightarrow y_3$, we have that $y_2$ and $y_3$ are not statistically independent, contrary to the implication of the causal Markov condition.

The fact the imposing implementation neutrality invalidates the causal Markov condition raises questions about the practice in the causation literature of taking the causal Markov condition as an axiom (as in Spirtes, Glymour and Scheines [14], for example). The fact that the causal Markov condition fails under a definition of causation that captures an important aspect of the ordinary-language meaning of causation—that proposed here—suggests that the causal Markov condition depends on factors not inherent in the idea of causation. We have not been informed as to what these conditions

---

\[8\]In the passage discussed in note 5, Hausman [7] observed that common parents of a variable are “probabilistically independent”. However, we noted that by this he meant that they are not causally connected. Whether probabilistic independence in this sense is equivalent to statistical independence depends on whether the causal Markov condition is satisfied. As noted below, it is not satisfied under implementation-neutral causality. Thus the result here is different from Hausman’s.
2.1 Example

Consider the following model:

\[ y_1 = x_1 + x_2 \]
\[ y_2 = x_3 + x_4 \]
\[ y_3 = x_1 + x_2 + x_3 + x_4 \]
\[ y_4 = x_1 + x_2 - x_3 - x_4 \]

(note that here we have supplied specific coefficient values as well as external sets). The causal form of this model is

\[ y_1 \leftarrow x_1 + x_2 \]
\[ y_2 \leftarrow x_3 + x_4 \]
\[ y_3 \leftarrow y_1 + y_2 \]
\[ y_4 \leftarrow y_1 - y_2, \]

with Figure 8 as its causal diagram. For simplicity we will assume that the \( x \) variables are normally as well as independently distributed, so as to be able to identify uncorrelatedness and independence. Here \( y_1 \) and \( y_2 \) are statistically independent due to the fact that their external sets are disjoint. We have that \( y_1 \) and \( y_2 \) are parents of \( y_3 \) (and \( y_4 \)), so the result illustrates
the general fact noted above that if any internal variable has more than one parent, either these are independent or one parent is a cause of the other.

Incidentally, the independence result can be generated via an indirect route using the graphical techniques developed by Pearl and others for analysis of causation in settings where implementation neutrality is not imposed. Specifically, there exist two paths from $y_1$ to $y_2$, but both are “blocked” by existence of the “colliders” (a variable is a collider if both arrows connecting it to the other variables of the path are inbound) $y_3$ and $y_4$. Therefore these paths do not transmit correlation. Independence of $y_1$ and $y_2$ results. Note that here the graphical analysis applies by virtue of the assumption that the external variables are independently distributed, not because of the causal Markov condition, which as we have seen does not generally apply under implementation neutrality. The result suggests that even though the conditions for causation analyzed here are different from those in the received literature, at least some of the graphical techniques for analysis of causation carry over.

The independence result does not extend to the children $y_3$ and $y_4$, however, except in special situations. For example, if the $x_i$ all have the same variance independence obtains. If $x_1$ and $x_2$ have higher (lower) variance than $x_3$ and $x_4$, then $y_3$ and $y_4$ will be positively (negatively) correlated.

3 Comparison with Implementation-Specific Causation

The graphical construction outlined in this paper differs in major respects from what is found in the literature on causation. The differences arise from the fact that causation here is associated with implementation neutrality, whereas elsewhere causation and implementation neutrality are not connected. Most important, interventions here consist of hypothetical alterations in the assumed values of external variables. In contrast, as noted above the usual treatment in the literature (based on Strotz and Wold [15]) involves modeling policy interventions on, say, $y_1$ by deleting from the model the equation determining $y_1$ and replacing it with the specification that $y_1$ is external.

There are several major problems with the Strotz-Wold procedure. It applies only in recursive systems, since in the presence of simultaneity $y_1$ is determined jointly with other variables in a group of several or many equa-
tions. Therefore there does not exist any obvious way to identify which equation is to be deleted. This aside, there is a troubling inconsistency between how solution values of variables are generated in the routine operation of the model—via realizations of external variables—and how they are modeled under a policy intervention—via suppressing equations and relabeling internal variables as external. What is it about policy interventions that motivates this difference? We are not told. Finally, altering the equations of a model appears to conflict with the basic idea of structuralness, which consists of invariance of the model structure under intervention.

The Strotz and Wold procedure assumes that causal models are modular, meaning that causal relations can be modified individually without invalidating the other equations of the model (modularity has been discussed widely in the causation literature; see, for example, Cartwright [1] and the works cited there). Under our treatment, in contrast, the existence or nonexistence of causal effects is calculated from the solution form, not the structural equations. The assumption that the structural model is recursive gives no information as to whether the appropriate restrictions are satisfied. If they are, it will be possible to write a set of structural equations in recursive form with the coefficients representing causation. If the restrictions are not satisfied one can still write structural equations in recursive form—this is always possible by implementing appropriate algebraic operations—but the resulting coefficients will not be interpretable as measuring causal relations.

Modeling interventions by respecifying internal variables as external—the Strotz-Wold procedure—implies that causation will be treated as if it were implementation neutral whether or not the model is such as to justify this treatment. If not, coefficients will be interpreted as causal when they do not support that interpretation. It is far from clear why one would want to take this route. The question “What is the effect of \( y_1 \) on \( y_2 \)?” will generally depend on what brings about the change in \( y_1 \), and the model encodes exactly this information in the equations determining \( y_1 \). Therefore the analyst can determine whether the question of causation has an unambiguous answer. Strotz and Wold and other proponents of analyzing interventions by reclassifying internal variables as external do not explain why it makes sense to adopt a framework that implies suppression of the question whether causation is unambiguous.

We observed that even when the Strotz-Wold device is rejected in favor of modeling interventions as alterations of variables that are explicitly identified as external, as we have done and as is sometimes done elsewhere in the
literature, the received treatment runs into difficulties. The problem, again,
is that most such exercises interpret structural coefficients as capturing causation without considering whether or not interventions are implementation neutral. This involves interpreting structural coefficients as causal when, as we have argued is always that case if implementation neutrality fails, the model does not support that interpretation.\footnote{A different approach to characterizing causal orderings can be found in the received literature (for example, Pearl \cite{13}). It is based on the observation that the assumed joint probability distribution of the model's variables can be factored into the product of conditional distributions. If the variables are arbitrarily ordered as $y_1, y_2, \ldots, y_n$, then the distribution $f$ can be written as

$$f(y_1, \ldots, y_n) = \Pi_j f(y_j | y_1, \ldots, y_{j-1}).$$

\text{(12)}$$

For each $j$ the conditioning set can be replaced by a minimal set consisting of the causal parents of $y_j$, such that the conditional distribution of $y_j$ depends on the parent(s) but not the other elements of the conditioning set.

The problem here is that the causal ordering so derived depends critically on the ordering of the variables, which is arbitrary. This is easy to verify from simple examples. Therefore the factorization (12) does not in fact identify a causal ordering.}

4 Application: Granger Causation\footnote{This section draws heavily on Cooley and LeRoy \cite{2}, although some of the discussion there is altered to accommodate the treatment here of causality.}

Granger \cite{5} proposed a definition of causation that can be implemented empirically without relying on theoretical restrictions: a time series $y_1 = \{y_{1t}\}$ \textit{Granger-causes} another time series $y_2$ if the optimal prediction of future values of $y_2$ based on past values of $y_2$ alone can be improved by including current and lagged values of $y_1$ as explanatory variables. It is asserted that if $y_1$ does not Granger-cause $y_2$, then $y_{2t}$ can be treated as strictly exogenous with respect to $y_{1t}$, so that correlations between the two can be interpreted as reflecting the causal effect of $y_2$ on $y_1$. The problem here is to determine the relation between Granger-causation and causation as defined in this paper.

Analysts recognized immediately that Granger-causation is not the same as causation as that term is used in ordinary discussion. For example, Granger pointed out that under the definition just stated cattle stamping their hooves before an earthquake implies that the cattle Granger-cause the
earthquake. Granger termed such cases “spurious causation”, implying that the question of how to define non-spurious causation remained open.

We need to understand the relation between Granger causation and causation as modeled here. Following Cooley-LeRoy [2], we formulate a two-variable vector autoregression generating the values of the money stock $m$ and gross domestic product $y$ (note that in this section we are using $y$ to denote GDP, not as representing a general internal variable):

\begin{align}
    m_t &= a_{my}y_t + b_{ym}m_{t-1} + b_{my}y_{t-1} + x_{1t}, \\
    y_t &= a_{ym}m_t + b_{ym}m_{t-1} + b_{yy}y_{t-1} + x_{2t}.
\end{align}

(13)

(14)

Here the external variables $x_{1t}$ and $x_{2t}$ are distributed independently, and are independent over time. The reduced form corresponding to this system is

\begin{align}
    m_t &= c_{mm}m_{t-1} + c_{my}y_{t-1} + u_{1t}, \\
    y_t &= c_{ym}m_{t-1} + c_{yy}y_{t-1} + u_{2t}.
\end{align}

(15)

(16)

GDP fails to Granger-cause the money stock if

\[ c_{my} = \frac{a_{my}b_{yy} + b_{my}}{1 - a_{my}a_{ym}} = 0. \]

(17)

The money stock is strictly exogenous with respect to GDP if $a_{my} = b_{my} = 0$. Strict exogeneity implies that GDP shocks do not feed back into the equation determining money, either currently or with a lag. Granger non-causation is a necessary condition for strict exogeneity, but not a sufficient condition.

We are interested in determining what parameter restrictions imply $m_t \rightarrow y_t$ under the definition proposed here. It turns out that even strict exogeneity of $m$ is not a sufficient condition for interpreting the regression coefficient of $y_t$ on $m_t$ as the causal coefficient associated with $m_t \rightarrow y_t$. This is so because the lagged values of $x_1$—the external variables that determine $y_t$ through their effect on $m_t$—also affect $y_t$ via $m_{t-1}$. Therefore we have a failure of implementation neutrality: characterizing an intervention as a hypothesized change in $m_t$ does not give enough information about the intervention to determine the resulting change in $y_t$.

The situation is exactly analogous to that discussed in Subsections 1.1 and 1.2. As in the earlier discussion there exists a condition on the coefficients
under which implementation neutrality does obtain. To derive it, we have only to adapt the derivations above to the current context. First, assume strict exogeneity of \( m_t \) \((a_{my} = b_{my} = 0)\). Then derive the solution form of the model above, in which \( y_t \) and \( m_t \) are expressed as functions of the current and lagged values of \( x_1 \) and \( x_2 \). Implementation neutrality occurs when the coefficient of each lagged value of \( x_{1j} \) in the equation for \( y_t \) has the same ratio to the corresponding coefficient in the equation for \( m_t \), for all \( j = t, t-1, t-2, \ldots \). Unsurprisingly, the condition for this to occur turns out to be \( b_{ym} = 0 \), so that \( m_{t-1} \) drops out of the equation for \( y_t \). The conclusion is that strict exogeneity of \( m_t \) alone, although stronger than Granger causation, is not strong enough to imply \( m_t \rightarrow y_t \); for \( m_t \rightarrow y_t \) we need also \( b_{ym} = 0 \) so that the implementation neutrality condition is satisfied.

We see that to make the transition from Granger-noncausation to causation in the sense of this paper, one has to make two further assumptions. The first is that \( c_{12} = 0 \) must be taken to mean \( a_{my} = b_{my} = 0 \). Analysts aware of this distinction frequently state that \( c_{12} = 0 \) is consistent with \( \theta = b_{12} = 0 \), but then go on to treat “is consistent with” as having the same meaning as “implies”. Second, as we have just seen implementation neutrality requires that one rule out \( m_{t-1} \) as an argument in the equation for \( y_t \). The conclusion is Granger causation is a specialized—and, to be sure, a very useful—form of forecastability, but it cannot be directly interpreted as having anything to do with implementation-neutral causation.

It may be that we are being too narrow in trying to relate Granger-causation to causation between current values of \( m \) and \( y \). The definition of causation here relates a single cause variable and a single effect variable, whereas Granger causation involves the stochastic processes \( m \) and \( y \). The suggestion is that a more general notion of causation is required. If so, the task at hand for proponents of Granger causation would seem to be to propose a more general characterization of (true) causation and then relate Granger causation to that.

5 Conditioning on Internal Variables

The result in Section 2 that the coefficient associated with any causal relation is identified and can be estimated consistently using least squares depends critically on the underlying assumption that external variables are independently distributed. If the analyst sees only data for which some variable \( y_i \)
takes on a single value or lies in a certain region, the relevant distribution assumed for the external variables is that conditional on these restrictions, not the unconditional distribution. Unless \( y_i \) depends in a simple way on the variables in its external set, which it generally will not, the conditional distribution of the external variables will display statistical dependence even if the unconditional distribution of the external variables incorporates independence. This situation will not affect the causal ordering of the variables, but it does invalidate the result that the coefficients associated with the causal ordering can be estimated consistently by least squares. This is so because failure of independence in the external variables implies that the error term covaries with the explanatory variable in the relevant regression, inducing bias and inconsistency.

As an extreme case, suppose that the analyst only has data in which \( y_i \) takes on a single value, for some \( i \). Obviously the coefficient associated with \( y_i \rightarrow y_j \) or \( y_j \rightarrow y_i \) is not identified, there being no variation in either the cause variable or the effect variable under the assumed restriction. A more common case occurs when the data for \( y_i \) are truncated, as by \( y_i \geq 0 \). In that case the sample regression coefficient associated with \( y_j \rightarrow y_k \) is identified, but it is not a consistent estimate of the associated causal coefficient if either \( y_j \) or \( y_k \) has external sets that overlap with that of \( y_i \). This is so because if \( y_i \) is subject to a restriction like \( y_i \geq 0 \) the relevant distributions of the external variables in \( \mathcal{E}(y_i) \) are those conditional on \( y_i \geq 0 \).

A simple example (adopted from Elwert [3]) illustrates this. Suppose that movie actors become famous if they are good looking or can act well, or both. Assume, probably realistically, that being good looking and being a good actor are independently distributed. If the analyst has a data set consisting only of actors who are famous, then any actor in that set who is not good looking must be a good actor, since otherwise he would not be famous. Similarly, any famous actor who is not a good actor must be good looking. Thus in the data set of famous actors there will be a negative correlation between being good looking and being a good actor, even though by assumption there is no such correlation in the general population. Any statistical exercise that requires that these variables be independent will be biased.

We will not discuss statistical procedures to deal with this problem since they do not deal directly with causal issues; the point here is only to demonstrate that the attractive statistical properties of least squares in estimating causal coefficients do not apply universally.
6 Conclusion

In this paper we propose a specific meaning for causation: one variable causes another if the effect of all interventions that produce a given alteration on the cause variable induce the same alteration on the effect variable. If this condition is satisfied the answer to the question “What is the effect of a change in \( y_1 \) on \( y_2 \)?” does not depend on what caused the assumed change in \( y_1 \). This, it seems to us, captures what scientists want to know when they investigate questions dealing with causation.

Philosophers sometimes reject limiting use of the term ‘causation’ to settings in which interventions are implementation-neutral. For example, Cartwright [1] states that “[w]e must be careful ... not to be misled by [LeRoy’s] own use of the language of ‘causal order’ to suppose it tells us whether and how much one quantity causally contributes to another” (p. 246). Why are we misled by this supposition? How much one quantity causally contributes to another is exactly what causation as defined here tells us, and is exactly what we want to know. And what meaning can we attach to a purported measure of the effects of an intervention on an internal variable if the model is such that the intervention is not implementation-neutral? In that case there is no alternative to redirecting the analysis to implementation-specific interventions on the external variables.

We saw that the algorithm proposed in this paper can be implemented to produce a directed acyclic graph in any equation system, not just in recursive systems as with the received graphical apparatus. In some settings the two types of graphs have different interpretations. Under the construction here systems with large simultaneous-equations blocks will have few or no causal arrows, reflecting the fact that in such models interventions usually cannot be adequately characterized by their effects on any one internal variable. The fact that causal orderings in nearly completely simultaneous models are nearly completely empty is not a flaw of the conception of causation developed here; rather, it is simple acknowledgement of the fact—familiar to economists from elementary economic analysis—that in simultaneous systems internal variable cannot intelligibly be interpreted as causing each other.\(^{11}\) However, we noted in several places above that causal graphs

\(^{11}\)All of this is in sharp contrast to the case under the received construction, in which simultaneity produces graphs that are cyclic. Thus our algorithm avoids the implication of the received construction that in nonrecursive models internal variables cause themselves, albeit indirectly, which seems to be an unfortunate implication.
can sometimes be used in the same way under our definition of causation as under other definitions. It would be worthwhile conducting a systematic comparison of inference from causal graphs in the two cases.

There remains the question of how to implement the definition of causation proposed here. At a minimum, the analysis here can play the role of raising questions about discussions of causation when there is no attempt to justify the implicit assumption of intervention neutrality. The idea is not to invalidate attributions of causation that do not include an explicit account of such an analysis, but rather to encourage clear communication about what exactly is involved in causal assertions.

References


