Causality and probability

1. Introduction

In *Physics* 2. 3, 194b16 – 195a3, Aristotle famously distinguishes four types of causes (material, formal, efficient, and final). Many economists deal with material causes: with the ultimate constituents of the economy (economic agents, their preferences, their behavior and so on). Some of them are interested in formal causes (the norms, institutions, or systemic properties that appear to determine the behavior of economic agents at least partly), and some of them (e.g. economists of the Marxist type) in final causes (in certain higher purposes of economic development). But perhaps the majority of economists are concerned with efficient causes: with the causes of shortages of donated organs, of an increase in inflation, of global inequality and so on.

When investigating efficient causes, economists stand in one of roughly two traditions: in the tradition of understanding efficient causes as raising the probability of their effects, or in the tradition of understanding them as causally dependent on an instrumental variable (or "instrument"), i.e. on a variable, on which not only the putative cause causally depends but also the putative effect (via the putative cause), and which doesn't causally depend on the putative effect or any other variable on which the putative effect causally depends. While the first tradition goes back to Hume, the second tradition has its roots in some of the work of the early econometricians (Haavelmo, Simon). The second tradition is younger than the first; but unlike the first tradition, the second tradition is at least compatible with Aristotelian approaches to efficient causation: with approaches that involve firm ontological commitments to powers, tendencies, or capacities.[1]

The present entry will be concerned with the first tradition almost exclusively (and only touch upon the second tradition in sections 6 and 7). It will briefly present and discuss the probability theories of causality of Suppes and Granger (sections 2 and 3) and introduce Zellner's idea of using causal laws to decide about the relevance of the variables and lags to be included in a model representing relations of Granger causality (section 4). It will then present and discuss causal Bayes nets theory (section 5) and emphasize that knowledge of causes that raise the probability of their effects can be employed for purposes of prediction, but less so for purposes of policy analysis (section 6). It will finally mention a number of

[1] Hoover (2001, p. 100), for instance, stands in the second tradition and characterizes his "structural account" of causality as "not inconsistent" with Cartwright's account of causes as capacities. Cartwright is sympathetic to probability theories of causality, but holds that (high) conditional probabilities (or regularities) only manifest "nomological machines", where a nomological machine is "a fixed (enough) arrangement of components, or factors, with stable (enough) capacities that in the right sort of stable (enough) environment, give rise to the kind of regular behavior that we represent in our scientific laws" (Cartwright, 1999, p. 50).

problems that are potentially inherent to attempts to infer causality in the sense of the second tradition from probabilities (section 7). In the remainder, ‘causation’ is taken to be synonymous with ‘efficient causation’.

2. Suppes on genuine causation

While Hume required constant conjunction of cause and effect, probability approaches to causality are content to understand causes as raising the probability of their effects. They say that $X = x$ causes $Y = y$ if the conditional probability of $Y = y$ given $X = x$ is greater than the unconditional probability of $Y = y$, formally: $P(Y = y | X = x) > P(Y = y)$, where X, Y … are random variables, i.e. functions from a sample or state space to a set into a range of values, where lower-case letters x, y … denote the values that X, Y … can take, and where P is a probability measure over the power set of that sample space, i.e. a function from that power set into the set of real numbers such that the Kolmogorov axioms are satisfied.

The power set of the sample space may also be understood as the set of propositions saying that $X = x$, $Y = y$ … Instead of propositions probability approaches to causality usually speak of events: of the event A of X attaining the value x, of the event B of Y attaining the value y … Suppes (1970, p. 12) interprets events as “instantaneous”, i.e. as occurring at a particular point in time; and he includes their time of occurrence in their formal characterization. So for him, ‘$P(A_t)$’ refers to the probability of the event A occurring at time t, where ‘A occurs at time t’ may mean as much as ‘X attains value x at time t’. Suppes (1970, p. 24) understands “cause” as “genuine cause” and defines “genuine cause” as “prima facie cause that is not spurious”. Thus, in order to understand his definition of “cause”, one needs to understand his definitions of “prima facie cause” and “spurious cause”.

His definition of “prima facie cause” runs as follows (cf. Suppes, 1970, p. 12):
(C_{PF}) $B_{t'}$ is a *prima facie* cause of A_t iff (i) $t' < t$, (ii) $P(B_{t'}) > 0$ and (iii) $P(A_t | B_{t'}) > P(A_t)$.
Condition (iii) is the condition that causes increase the probability of their effect, and condition (ii) is needed because in the definition of conditional probability – $P(A_t | B_{t'}) = P(A_t \wedge B_{t'})/P(B_{t'})$ – $P(B_{t'})$ is the denominator, and because the denominator must not be equal to zero. Condition (i) implies that $B_{t'}$ occurs earlier than A_t in time. Why does Suppes introduce that condition? One answer is that the relation ‘… increases the probability of …’ is symmetric because $P(A_t | B_{t'}) > P(A_t)$ is equivalent to $P(B_{t'} | A_t) > P(B_{t'})$, that the relation ‘… causes …’ is asymmetric, and that temporality is capable of turning the relation ‘… increases the probability of …’ into an asymmetric one. A second answer is that the Humean tradition, in which Suppes stands, holds that causality is intrinsically linked to temporality.

His definition of “spurious cause” runs as follows (Suppes, 1970, pp. 21-2):

(C_S) $B_{t'}$ is a spurious cause of A_t iff $B_{t'}$ is a prima facie cause of A_t, and there is a $t'' < t'$ and an event $C_{t''}$ such that (i) $C_{t''}$ precedes $B_{t'}$, (ii) $P(B_{t'} \wedge C_{t''}) > 0$ and (iii) $P(A_t \mid B_{t'} \wedge C_{t''}) = P(A_t \mid C_{t''})$.

In other words: $B_{t'}$ is a spurious cause of A_t iff $B_{t'}$ is a prima facie cause of A_t, $C_{t''}$ precedes $B_{t'}$, and $C_{t''}$ "screens off" A_t from $B_{t'}$. The notion of a spurious cause is needed to rule out cases in which *prima facie* causes do not represent genuine causes. A falling barometer, for instance, is a *prima facie* cause but not a genuine cause of an upcoming storm. Atmospheric pressure that precedes both the falling barometer and the upcoming storm screens off the upcoming storm from the falling barometer.

(C_S) is not the only definition of spurious causation that Suppes (1970, pp. 21-8) brings into play, and besides "prima facie cause" and "spurious cause" he defines "direct cause", "sufficient cause" and "negative cause". But a consideration of these definitions lies beyond the purposes of this entry. More immediately relevant to these purposes is a consideration of the problems that Suppes' account of genuine causation faces, and that require specific solutions. These problems are of essentially two kinds; they both suggest that condition (iii) of (C_{PF}) cannot be a necessary condition for $A_{t'}$ causing B_t.

The first problem is that it seems that $B_{t'}$ can turn out to be a cause of A_t even though $P(A_t \mid B_{t'}) < P(A_t)$. This problem can be illustrated by an example that Suppes (1970, p. 41) himself discusses. The example is that of a golfer with moderate skill who makes a shot that hits a limb of a tree close to the green and is thereby deflected directly into the hole, for a spectacular birdie. If A_t is the event of making a birdie and $B_{t'}$ the earlier event of hitting the limb, we will say that $B_{t'}$ causes A_t. But we will also say that $P(A_t \mid B_{t'}) < P(A_t)$: that the probability of his making a birdie is low, and that the probability of his making a birdie, given that the ball hits the branch, is even lower.

Does the example show that condition (iii) of (C_{PF}) cannot qualify as a necessary condition for $B_{t'}$ causing A_t? Suppes (1970, p. 42-3) argues for a negative answer. He argues that definition (C_{PF}) can be defended if condition (iii) is relativized to background information $K_{t'}$:

(C_{PF}') $B_{t'}$ is a *prima facie* cause of A_t iff (i) $B_{t'} \wedge K_{t'}$ precedes A_t, (ii) $P(B_{t'} \wedge K_{t'}) > 0$ and (iii) $P(A_t \mid B_{t'} \wedge K_{t'}) > P(A_t \mid K_{t'})$.

Thus, if $K_{t'}$ is e.g. the event of the shot being deflected in a specific angle, then the probability of the golfer's making a birdie, given that the ball hits the branch and is deflected in a specific angle, might well be higher than the probability of his making a birdie. Suppes (1970, p. 42) adds that such relativization to background knowledge "can be useful, especially in theoretical contexts".

The second problem is a fact about probabilities that is well known to statisticians and is often referred to as "Simpson's paradox". The fact is that any association between two variables which holds in a given population – $P(Y = y|\ X = x) > P(Y = y)$, $P(Y = y|\ X = x) < P(Y = y)$ or $P(Y = y|\ X = x) = P(X = x)$) – can be reversed in a subpopulation by finding a third variable that is correlated with both. Consider, for instance, the population of all Germans. For the population of all Germans, the conditional probability of getting a heart disease, given that an individual smokes, is *higher* than the unconditional probability of getting a heart disease. But for a subpopulation of Germans, in which all smokers exercise, the conditional probability of getting a heart disease, given that an individual smokes, is *lower* than the unconditional probability of getting a heart disease – at least if exercising is more effective at preventing heart disease that smoking at causing it.

The fact itself is not a paradox. But Cartwright (1979, p. 421) points out that the paradox arises if we define causation of $Y = y$ by $X = x$ in terms of $P(Y = y|\ X = x) > P(Y = y)$. If we define causation of $Y = y$ by $X = x$ in terms of $P(Y = y|\ X = x) > P(Y = y)$, then causation of $Y = y$ by $X = x$ will depend on the population that we select when establishing $P(Y = y|\ X = x) > P(Y = y)$. At the same time, we have the strong intuition that causation should be independent of specific populations. Cartwright (1979, p. 423) proposes to dissolve the paradox by conditioning $Y = y$ on the set of "all alternative causal factors" of $Y = y$. Conditioning $Y = y$ on such a set would render Suppes' definition of genuine causation circular: causal vocabulary would show up in both the *definiendum* and the *definiens*. Interestingly, however, few philosophers hold that non-circularity is absolutely necessary.[2]

3. Granger causality

Perhaps the most influential explicit approach to causality in economics is that of Granger (1969; 1980). Like Suppes, Granger stands in the Humean tradition of understanding causes as raising the probability of their effect; and like Suppes, he believes that causality is intrinsically linked to temporality. But unlike Suppes, Granger (1980, p. 330, notation modified) defines 'causation' as a relation between variables:

(*GC*) X_t Granger-causes Y_{t+1} if and only if $P(Y_{t+1} = y_{t+1} \mid \Omega_t = \omega_t) \neq P(Y_{t+1} = y_{t+1} \mid \Omega_t = \omega_t - X_t = x_t)$,

where Ω_t is the infinite universe of variables dated t and earlier. The temporal ordering of X_t and Y_{t+1} guarantees that the relation between X_t and Y_{t+1} is asymmetric, and conditioning on

[2] Woodward (2003, pp. 104-5), for instance, makes it clear that he is interested in the conceptual entanglement between causation and intervention, and not in any non-circular definition or reductive account of causality. Similarly, Hoover (2001, p. 42) claims that "[circularity] is less troubling epistemologically than it might seem to be ontologically".

$\Omega_t = \omega_t$ immunizes (*GC*) against circularity, spuriousness and the problem that causes might lower the probability of their effects.

Spohn (2012, p. 442) points out that it is "literally meaningless and an abuse of language" to speak of variables themselves as causing other variables. 'X_t causes Y_{t+1}' may either mean '$X_t = x_t$ causes $Y_{t+1} = y_{t+1}$' or 'Y_{t+1} causally depends on X_t', where the causal dependence of Y_{t+1} on X_t is to be understood as a relation that obtains between X_t and Y_{t+1} if some event $X_t = x_t$ causes some event $Y_{t+1} = y_{t+1}$.[3] The context of time series econometrics suggests that (*GC*) is to be read in the first sense (cf. Spohn, 1983, pp. 85-6). The phrase 'X_t Granger-causes Y_{t+1}' will be retained in the remainder because economists and econometricians have become accustomed to its use. It should be kept in mind, however, that the phrase is to be understood as synonymous with '$X_t = x_t$ causes $Y_{t+1} = y_{t+1}$'.

Granger (1980, p. 336) points out himself that (*GC*) is not "operational" because practical implementations cannot cope with an infinite number of variables with an infinite number of lags. But econometricians think that in order to test for "Granger causality", they need to select only the relevant variables and only the relevant number of lags. Sims (1972), for instance, uses two variables (for money and GNP) and 4 future and 8 past lags to show that money Granger-causes GNP, and not the other way around. Later, as part of a general critique of the practice of using *a priori* theory to identify instrumental variables, Sims (1980a) advocates vector autoregression (VAR), which generalizes Granger causality to the multivariate case. The following two-equation model, for instance, is a VAR model for two variables and one lag:

(1) $y_{t+1} = \alpha_{11}y_t + \alpha_{12}x_t + \varepsilon_{1t+1}$,

(2) $x_{t+1} = \alpha_{21}y_t + \alpha_{22}x_t + \varepsilon_{2t+1}$,

where the α_{ij} are parameters and the E_{it+1} random error terms. X_t is said to Granger-cause Y_{t+1} if $\alpha_{12} \neq 0$; and Y_t is said to Granger-cause X_{t+1} if $\alpha_{21} \neq 0$.

While definition (*GC*) avoids some of the problems that competing definitions face (circularity, spuriousness, the problem of causes that lower the probability of their effects), objections have been raised to implementations of (*GC*), i.e. to empirical procedures of testing for Granger causality. Hoover (1993, pp. 700-705) lists three problems that stand in the way of a temporal ordering of cause and effect in macroeconomics. The first problem is that in macroeconomics, it is difficult to rule out contemporaneous causality because data are

[3] Instead of relations of 'causal dependence' theorists sometimes speak of relations of 'type-level causation'. Both ways of speaking refer to relations between variables (e.g. to the relation between income and consumption in general), and not to relations between events (i.e. not to relations like that between the event of US income attaining a specific value in Q4 2019 and the event of US consumption attaining a specific value in Q4 2019).

reported most often annually or quarterly. Hoover (1993, p. 702) cites Granger as suggesting that contemporaneous causality could be ruled out if data were sampled at fine enough intervals. But Hoover (1993, p. 702) responds that "such finer and finer intervals would exacerbate certain conceptual difficulties in the foundations of economics"; and he cites GNP as an example: "There are hours during the day when there is no production; does GNP fall to nought in those hours and grow astronomically when production resumes? Such wild fluctuations in GNP are economically meaningless."

The second problem is that there are hidden variables that (like expectations) cannot be included among the regressors in VAR models even though they are likely to be causally relevant. And the third problem is that economic theory (no matter which) provides reasonably persuasive accounts of steady-states, i.e. of hypothetical economic configurations that feature constant rates, quantities, and growth rates, and that are timeless in the sense that they result if time is allowed to run on to infinity. Hoover (1993, p. 705) admits that proponents of Granger causality might respond that "if macroeconomics cannot be beat into that mold [of temporal ordering], so much the worse for macroeconomics". But Hoover (1993, p. 706) also argues that in macroeconomics, causal questions like 'Will interest rates rise if the Fed sells $50M worth of treasury bonds?' are sensible and well formulated, and that our concepts of causality need to be suitable for their formulation and interpretation.

Another prominent objection that has been raised to Granger-causality tests says that it is impossible to select the relevant number of variables and lags without (explicit or implicit) reliance on economic theory or background knowledge. Sims's subsequent work on the relation of Granger-causality between money and GNP indicates why this objection is important. When he included four variables (money, GNP, domestic prices and nominal interest rates) and twelve past lags in a VAR model, the above-mentioned result of money Granger-causing GNP no longer obtained.[4] A relation of Granger-causality thus crucially depends on the number of variables and lags that are deemed to be relevant. And who is to decide about the relevance of variables and lags, and how?

4. Zellner on causal laws

Zellner (1979; 1988) can be read as responding to that question when defining 'causality' in terms of "predictability according to a law or set of laws".[5] He claims that laws may be deterministic or stochastic, qualitative or quantitative, micro or macro, resulting from

[4] The new result stated that money accounted for only 4% of the variance in GNP (cf. Sims 1980b).

[5] Zellner (1979, p. 12; 1988, p. 7) points out that he adopts that definition from Herbert Feigl.

controlled or uncontrolled experiments, involving simultaneous or non-simultaneous relations and so on, and that the only restrictions that need to be placed on laws relate to logical and mathematical consistency and to the ability to explain past data and experience and to predict future data and experience. While these restrictions are not "severe", they imply that statistical regressions or autoregressions cannot qualify as laws because "they generally do not provide understanding and explanation and often involve confusing association or correlation with causality" (Zellner 1988, p. 9). Similarly, theories cannot qualify as laws if they are "based on impossible experiments or on data that can never be produced" (Zellner 1988, p. 9).

Sketching a rudimentary theory of the psychology of scientific discovery, Zellner (1988, pp. 9-12) suggests that the discovery of laws proceeds in roughly three steps. In a first step, "the conscious and unconscious minds interact […] to produce ideas and combinations of ideas using as inputs at least (1) observed or known past data and experience, (2) a space of known theories, and (3) future knowable data and experience." In a second step, "the conscious mind […] decides the general nature or design of an investigation." This means that it selects a specific "phenomenon" from its pool of ideas or combinations of ideas, and that it develops "an appropriate theory or model that is capable of explaining the phenomenon under investigation and yielding predictions." With respect to the development of that theory or model, Zellner remarks that it requires "hard work, a breadth of empirical and theoretical knowledge, consideration of many possible combinations of ideas, luck, and a subtle interaction between the conscious and unconscious minds." He also argues that "focusing attention on sophisticatedly simple models and theories is worthwhile."

The third and final step is that of demonstrating that "the suggested model or theory actually does explain what it purports to explain by empirical investigations using appropriate data." That demonstration requires the frequent use of new data to test not only the model or theory itself, but also its implications, such as predictions about as yet unobserved phenomena. Whenever new data is used to test the model or theory or its implications successfully, the degree of reasonable belief or confidence in the model or theory increases. The degree of reasonable belief in the model or theory corresponds to the posterior probability that can be assigned to that model or theory and computed using Bayes' theorem: $P(H|E) = P(E|H) \cdot P(H) / P(E)$, where H is a proposition summarizing the model or theory and E an 'evidential proposition' referring to new data that can be used to test H. Zellner (1988, p. 16) says that "a theory can be termed a causal law" if the posterior probability that can be assigned to it is "very high, reflecting much outstanding and broad-ranging performance in explanation and prediction."

Zellner can be read as responding to the question of how to decide about the relevance of variables and lags because causal laws may include well-confirmed theories about the strength of the parameters that can be included in a VAR model. If that strength happens to be among the phenomena that the conscious mind decides to investigate, the mind aims to develop an appropriate model or theory *H* that is capable of quantifying and explaining that strength. Once *H* is developed, it can be subjected to a Bayesian updating procedure in which a prior probability is assigned to *H*, in which the likelihood of *E* given *H* is evaluated, and in which data *E* is collected to compute the posterior probability of *H* in accordance with Bayes' theorem. The posterior probability of *H* then serves as its prior probability when new data *E* is collected to test *H* (or any of its implications) again. Once the posterior probability of *H* is "very high", *H* can be viewed as a causal law that supports decisions about the relevance of the variables and lags to be included in a VAR model: the greater the strength of a parameter, the more relevant its corresponding (lagged) variable.

Zellner's definition of "causality" combines with his rudimentary theory of the psychology of scientific discovery to imply an interesting response to the question of how to decide about the relevance of variables and lags. But problems pertain to the Bayesian updating procedure that marks the third of the three steps that scientific discovery takes according to his theory.[6] And even if his theory were accurate, the question arises whether it doesn't just make manifest how entirely difficult it is decide about the relevance of the variables and lags to be included in a VAR model. Zellner (1988, pp. 17-19) cites Friedman's theory of the consumption function as a theory with a very high posterior probability. But the very high posterior probability of that theory may well be exceptional.

5. Causal Bayes nets theory

One final probability approach to causality is causal Bayes nets theory. Causal Bayes nets theory was first developed outside economics (substantially foreshadowed in Spohn 1980, and then developed in detail by Spirtes, Glymour, and Scheines 1993 and Pearl 2000), but has been applied in economics and econometrics soon after (Bessler and Lee 2002, Demiralp and Hoover 2003). Unlike the approaches of Suppes and Granger, causal Bayes nets theory is primarily interested in relations of causal dependence and analyzes causal relations irrespectively of any temporal ordering. It consequently focuses on relations of *direct* causal dependence more explicitly.

[6] Cf. Norton (2011) for a particularly concise and thorough discussion of these problems.

At the center of causal Bayes nets theory[7] is the notion of a directed, acyclic graph (DAG). A DAG is a tuple $\langle \rightarrow, \mathbf{V} \rangle$, where $\mathbf{V}$ is a non-empty finite set of pre-selected variables and $\rightarrow$ an acyclic relation on $\mathbf{V}$: there are no variables $X, \ldots, Y \in \mathbf{V}$ such that $X \rightarrow \ldots \rightarrow Y$ and $Y \rightarrow X$. A DAG is a *causal* graph if the arrow $\rightarrow$ can be interpreted as representing a relation of direct causal dependence between the variables in $\mathbf{V}$. In order to understand the notion of direct causal dependence that is involved here, one needs to become acquainted with a bit of graph theoretical notation and the three axioms that determine the relation between causality and probability according to causal Bayes nets theory.

Consider the graph theoretical notation first: In a DAG, $X \in \mathbf{V}$ is said to be

- a *parent* of $Y \in \mathbf{V}$ if and only if $X \longrightarrow Y$ (the set of parents of Y is denoted by $\mathbf{pa}(Y)$).
- a *child* of $Y \in \mathbf{V}$ if and only if Y is a parent of X.
- an *ancestor* of $Y \in \mathbf{V}$ if and only if there are $X, \ldots, Y \in \mathbf{V}$ such that $X \longrightarrow \ldots \longrightarrow Y$ (the set of ancestors of Y is denoted by $\mathbf{an}(\mathrm{Y})$).
- a *descendant* of $Y \in \mathbf{V}$ if and only if Y is an ancestor of X.
- a *non-descendant* of $Y \in \mathbf{V}$ if and only $X \neq Y$ and X is not a descendant of Y (the set of non-descendants of Y is denoted by $\mathbf{nd}(Y)$).

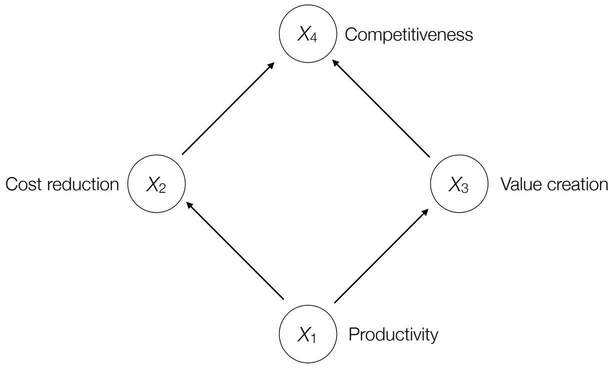
Now turn to the three axioms (cf. Spirtes, Glymour, and Scheines 1993, pp. 29-32). Let $\langle \rightarrow, \mathbf{V} \rangle$ be a causal graph and P a probability measure over the power set of the sample space, and let $\perp_P$ stand for probabilistic independence. Then P satisfies the so-called

- Causal Markov Condition if and only if for all $X \in \mathbf{V}$ $X \perp_P \mathbf{nd}(X) - \mathbf{pa}(X) \,/\, \mathbf{pa}(X)$.
- Causal Minimality Condition if and only if for all $X \in \mathbf{V}$ $\mathbf{pa}(X)$ is the smallest subset of variable set $\mathbf{Y}$ such that $X \perp_P \mathbf{nd}(X) - \mathbf{Y} \,/\, \mathbf{Y}$.
- Causal Faithfulness Condition if and only if for all subsets $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ of $\mathbf{V}$ $\mathbf{X} \perp_P \mathbf{Y} \,/\, \mathbf{Z}$ holds only if $\mathbf{X} \perp_P \mathbf{Y} \,/\, \mathbf{Z}$ is entailed by P's satisfaction of the causal Markov and minimality conditions.

Informally, the causal Markov condition says that the parents of X screen off X from all other non-descendants of X. The causal minimality condition says that P would no longer satisfy the causal Markov condition if any of the parents of X were excluded from $\mathbf{pa}(X)$; it requires that there be exactly one minimal set of parents of X that screens off X from all its other non-descendants. Finally, the faithfulness condition says that there are no accidental conditional independencies: that all the conditional independencies that the causal Markov and minimality conditions make reference to reflect relations of causal dependence.

[7] Much of the notation that the present section uses to describe causal Bayes nets theory is borrowed from Spohn (2012: section 14.8).

If P satisfies the causal Markov, minimality and faithfulness conditions and $\langle\rightarrow, \mathbf{V}\rangle$ is a causal graph, then P combines with $\langle\rightarrow, \mathbf{V}\rangle$ to form a causal Bayes net. As an example of a causal Bayes net, consider the following graph



and imagine that you are worried about the competitiveness (X_4) of your firm, and that you ponder about it in terms of productivity (X_1), cost reduction (X_2) and value creation (X_3) and the probabilistic independencies that you think obtain between these variables. Then the causal Markov condition entails that $X_2 \perp_P X_3 \ / \ X_1$ (that X_2 and X_3 are probabilistically independent, given their common cause X_1), and that $X_4 \perp_P X_1 \ / \ \{X_2, X_3\}$ (that X_2 and X_3 screen off X_4 from X_1). The minimality condition entails that it is not the case that $X_2 \perp_P X_3$ (otherwise $\{X_1\}$ would not be the minimal set given which X_2 is independent of its non-descendant X_3, and vice versa), and that it is not the case that $X_4 \perp_P X_2 / X_3$ or $X_4 \perp_P X_3 / X_2$ (X_2 and X_3 must make a difference, given the other). Finally, the faithfulness condition requires that probabilistic dependencies do not disappear when there are causal chains: that it be not the case that $X_4 \perp_P X_1 / X_2$, $X_4 \perp_P X_1 / X_3$, or $X_4 \perp_P X_1$.

Spirtes, Glymour, and Scheines make it clear that they do not expect the three axioms to hold universally. They point out that the causal Markov condition might be violated in quantum physics, and that the causal faithfulness condition is violated on occasion (in the foregoing example it would be violated if $X_4 \perp_P X_1$ because the direct influences of X_2 and X_3 cancel out each other). But they also say of the three axioms that "their importance – if not their truth – is evidenced by the fact that nearly every statistical model with a causal significance we have come upon in the social scientific literature satisfies all three" (Spirtes, Glymour, and Scheines 1993, p. 53). What they could have stated more clearly is that in order to satisfy the three axioms, a statistical model or set of variables needs to be causally sufficient: it needs to include each proximate common cause of any two variables in **V**; otherwise the probabilistic independencies will inadequately reflect relations of direct causal dependence.

Spohn (2012, p. 501) points out that causal sufficiency might be difficult to achieve. The common causes of any two variables in **V** might go back as far as the big bang or simply slip off our radars, especially when they are hidden, i.e. non-measurable and causally relevant. Hoover (2001, p. 168) analyzes the repercussions of this difficulty for the case of economics. He argues that the faithfulness condition might be violated whenever expectations operate because expectations are hidden and take positions in causal relations that might fail to be reflected by conditional independencies. Hoover (2001, p. 167) argues, moreover, that macroeconomics poses "systematic threats to the Causal Markov Condition" because the "search for an unmeasured conditioning variable may end in the crossing of the micro/macro border before an appropriate conditioning variable could be located".

Spirtes, Glymour, and Scheines refrain from defining 'causation' explicitly and prefer to understand conditional independencies as reflecting relations of direct causal dependence. Spohn (2012, pp. 508-509), by contrast, proposes to define direct causal dependence in terms of the conditional independencies. He proposes, more specifically, to say that Y causally depends on X directly if and only if not $Y \perp_P X \,/\, \mathbf{nd}(Y) - X$, i.e. if and only if it is not the case that Y is probabilistically independent of X, given all the non-descendants of Y except X. He emphasizes that this definition is problematic because it relativizes the notion of direct causal dependence to that of a set **V** of pre-selected variables: change that set, and you will change the conditional independencies, and with them relations of direct causation. But he also suggests that the problem can be solved by de-relativizing the notion of direct causal dependence, i.e. by defining it for a "universal frame" or universal set of variables.

6. Policy or prediction?

Sections 3 and 4 called attention to the problem that economists cannot establish the claim that X_t Granger-causes Y_{t+1} unless they manage to include in a VAR model only the relevant number of variables and lags. Assume that despite this problem, they manage to establish the claim that X_t Granger-causes Y_{t+1}. Can they now predict the value that Y_{t+1} is going to attain if they know the value of X_t? Can they predict, for instance, the value that GNP will attain in t+1 if they know the value that money takes in t? Most economists believe that the answer is positive. Granger (1969, p. 428) suggests that prediction is in fact the principal purpose of searching for relations of Granger causality. And in statistics, there are standard procedures for computing the expected value of Y_{t+1} when the values of the other variables and lags in the model are given. Economists might not be able to predict the exact value of Y_{t+1} (e.g. GNP), but they can state the probability with which Y_{t+1} can be expected to attain a specific value.

An entirely different question is whether economists can predict the value that Y_{t+1} *would* attain *if* they were to *control* X_t to a specific value. Assume again that they know that X_t (standing e.g. for money) Granger-causes Y_{t+1} (standing e.g. for GNP); does that imply that they know the value that Y_{t+1} would attain if they managed to set X_t to a specific value? Most economists agree that the answer is negative. In order to see that the answer is negative, consider again equations (1) and (2) of section 3. In order to be able to predict the value that Y is going to attain in t+1, one needs to condition equation (1) on the observations of X and Y in t and take the expectation of Y_{t+1}:

(3) $E(Y_{t+1}| \, y_t, x_t) = \alpha_{11}y_t + \alpha_{12}x_t + E(\mathrm{E}_{1t+1}| \, y_t, x_t)$.

But in order to be able to predict the value that Y_{t+1} would attain if X_t were controlled to x_t, one would need to condition equation (1) on the observations of X and Y in t and take the expectation of a *counterfactual* quantity. The expectation of that quantity is calculated in the same way as in (3). But in order to understand that quantity as counterfactual, one would need to understand the relation between X_t and Y_{t+1} as causal in the sense of the second tradition mentioned in the introduction: one would need to assume that there is an instrumental variable I_t (standing e.g. for the federal funds rate) that causes X_t, that causes Y_{t+1} only via X_t, and that isn't caused by E_{1t+1}; one would need to interpret equation (1) as a structural equation (and not as a regression equation) and E_{1t+1} as encompassing omitted variables that cause Y_{t+1} (and not as a regression error).[8]

Thus knowledge that X_t Granger-causes Y_{t+1} is not sufficient for (does not imply) knowledge of the value that Y_{t+1} would take if X_t were controlled to x_t. Might one perhaps say that knowledge that X_t Granger-causes Y_{t+1} is *necessary* for knowledge of the value that Y_{t+1} would take if X_t were controlled to x_t? Unfortunately, the answer is still negative. In order to see that the answer is negative, consider the following model of structural equations:

(4) $y_{t+1} = \theta x_{t+1} + \beta_{11}y_t + \beta_{12}x_t + v_{1t+1}$,

(5) $x_{t+1} = \gamma y_{t+1} + \beta_{21}y_t + \beta_{22}x_t + v_{2t+1}$,

where θ, γ and the β_{ij} represent parameters and the N_{it+1} structural errors, i.e. errors encompassing omitted variables that are causally relevant. Solving the current values out of these equations yields the reduced form equations, which coincide with equations (1) and (2) such that $\alpha_{11} = (\beta_{11} + \theta\beta_{21})/(1 - \theta\gamma)$, $\alpha_{12} = (\beta_{12} + \theta\beta_{22})/(1 - \theta\gamma)$, $\alpha_{21} = (\gamma\beta_{11} + \beta_{21})/(1 - \theta\gamma)$, $\alpha_{22} = (\gamma\beta_{12} + \beta_{22})/(1 - \theta\gamma)$, $\varepsilon_{1t} = (v_{1t} + \theta v_{2t})/(1 - \theta\gamma)$, $\varepsilon_{2t} = (\gamma v_{1t} + \theta v_{2t})(1 - \theta\gamma)$. In order for Granger causality (or knowledge thereof) to qualify as a necessary condition of causality in the sense of the second tradition (or knowledge thereof), α_{12} in equation (1) would need to be unequal to zero. But Jacobs, Leamer, and Ward (1979, pp. 402-5) show (for a similar model) that

[8] One would need to interpret E_{1t+1}, more specifically, as encompassing omitted variables that adopt certain values and cause Y_{t+1} in t+1 or earlier.

there are cases in which α_{12} is equal to zero: cases in which e.g. $\beta_{12} = -\theta\beta_{22}$. And Hoover (2001, pp. 152-3) points out that these cases are not among the exotic ones that economists can neglect with a clear conscience.

While the question relating to the value that Y_{t+1} is going to attain if the value of X_t is reported to be x_t arises in contexts of forecasting, the question relating to the value that Y_{t+1} would attain if X_t were set to x_t by intervention arises in contexts of policy analysis. It goes without saying that complementing knowledge of causality in the sense of the second tradition with knowledge of Granger causality is likely to be helpful in both contexts. And perhaps knowledge of causality in the sense of the second tradition yields better predictions than knowledge of Granger causality (cf. Pearl, 2000, p. 31). But the decisive point of the foregoing analysis is that policy analysis requires knowledge of causality in the sense of the second tradition.

Many economists believe that policy analysis is the ultimate justification for the study of economics (cf. e.g. Hoover, 2001, p. 1); and that belief might explain why some of them hold that only the second tradition deals with causality in the strict sense of the term. Sargent (1977, p. 216), for instance, states that "Granger's definition of a causal relation does *not*, in general, coincide with the economists' usual definition of one: namely, a relation that is invariant to interventions in the form of imposed changes in the processes governing the causal variables." In econometric textbook expositions of the concept of causality, one is likewise likely to find the observation that "Granger causality is not causality as it is usually understood" (Maddala and Lahiri, 2009, p. 390).

7. Common effects and common causes

The result of the preceding section has been that (knowledge of) Granger causality is neither a necessary nor sufficient condition of (knowledge of) causality in the sense of the second tradition. Does that result generalize to the claim that (knowledge of) causality in the sense of the second tradition can *never* be inferred from (knowledge of) probabilities? Hoover (2009, p. 501) defends a negative answer when suggesting that "some causal claims may be supported by facts about probability models that do not depend on assumptions about the truth of these very same causal claims." The causal claims that he discusses include the claim that Z causally depends on both X and Y and the claim that X and Y causally depend on Z. The primary purpose of the present and final section is to point to the problems that are potentially inherent to attempts to infer these claims from probability models.

It is, of course, impossible to observe the relations of causal dependence that might (or might not) obtain between X, Y, and Z directly. But one might be able to observe realizations of X, *Y* and *Z*. And Hoover thinks that it is possible to specify an adequate probability model for these realizations independently of any assumptions about the causal relations that might (or might not) obtain between *X*, *Y*, and *Z*. In some of his work, Hoover (2001, pp. 214-7) advocates the application of LSE methodology to specify adequate probability models. LSE methodology operates by (i) specifying a deliberately overfitting general model, by (ii) subjecting the general model to a battery of diagnostic (or misspecification) tests (i.e. tests for normality of residuals, absence of autocorrelation, absence of heteroscedasticity and stability of coefficients), by (iii) testing for various restrictions (in particular, for the restriction that a set of coefficients is equal to the zero vector) in order to simplify the general model, and by (iv) subjecting the simplified model to a battery of diagnostic tests. If the simplified model passes these tests, LSE methodology continues by repeating steps (i) – (iv), i.e. by using the simplified model as a general model, by subjecting that model to a battery of diagnostic tests etc. Simplification is complete if any further simplification either fails any of the diagnostic tests or turns out to be statistically invalid as a restriction of the more general model.

Let us assume that the application of LSE methodology has resulted in the following normal model of *X*, *Y* and *Z* (cf. Hoover, 2009, p. 502, notation modified):

$$(X, Y, Z) \sim N\,(\mu_X, \mu_Y, \mu_Z, \sigma^2{}_X, \sigma^2{}_Y, \sigma^2{}_Z, \rho_{XY}, \rho_{XZ}, \rho_{YZ}),$$

where μ_X, μ_Y and μ_Z are the three means, $\sigma^2{}_X$, $\sigma^2{}_Y$ and $\sigma^2{}_Z$ the three variances, and ρ_{XY}, ρ_{XZ} and ρ_{YZ} the three covariances or population correlations of the model. Hoover argues that the model supports the claim that *Z* causally depends on both *X* and *Y* if it satisfies the antecedent of the common effect principle, and that it supports the claim *X* and *Y* causally depend on *Z* if it satisfies the antecedent of the common cause principle. The two principles can be restated as follows (cf. Hoover 2009):

- Principle of the Common Effect: If *X* and *Y* are probabilistically independent conditional on some set of variables (possibly the null set) excluding *Z*, but are probabilistically dependent conditional on *Z*, then *Z* causally depends on both *X* and *Y* (then *Z* forms an unshielded collider on the path *XZY*).
- Principle of the Common Cause: If *X* and *Y* are probabilistically dependent conditional on some set of variables (possibly the null set) excluding *Z*, but are probabilistically independent conditional on *Z*, then *X* and *Y* causally depend on *Z*.

Hoover argues, more specifically, that the normal model of *X*, *Y* and *Z* supports the claim that *Z* causally depends on both *X* and *Y* if $\rho_{XY} = 0$ and $\rho_{XY|Z} \neq 0$, and that it supports the claim that *X* and *Y* causally depend on *Z* if $\rho_{XY} \neq 0$ and $\rho_{XY|Z} = 0$.

There are three problems that are potentially inherent to attempts to infer these claims from probability models. The first problem is that in practice, LSE methodology might be incapable of implementation without data mining, which Mayo (1996, pp. 316-7) characterizes as data use for double duty, i.e. as the use of data to arrive at a claim (e.g. at the claim that *Z* causally depends on both *X* and *Y*, or that *X* and *Y* causally depend on *Z*) in such a way that the claim is constrained to satisfy some criteria (e.g. absence of misspecification), and that the same data is regarded as supplying evidence in support of the claim arrived at. Spanos (2000), however, argues that there are problematic and non-problematic cases of data mining.

The second problem is that Hoover's claim that an adequate probability model can be specified independently of any causal assumptions might not be accurate. If there are hidden variables (i.e. variables that cannot be measured and are known to be causally relevant), then these variables cannot be included in a deliberately overfitting general model, and then the model resulting from the application of LSE methodology cannot be said to be adequate.[9] But even if the probability model can be said to be adequate, there will be the third problem that neither principle obtains in cases in which $\rho_{XY} \neq 0$ denotes a nonsense correlation like that between higher than average sea levels and higher than average bread prices (Sober, 2001, p. 332), or that between cumulative rainfall in Scotland and inflation (Hendry, 1980, pp. 17-20). It would be absurd to ask for the variable, which causally depends on *X* and *Y*, or for the variable, on which *X* and *Y* causally depend if *X* and *Y* were correlated in a way that doesn't make any sense.

Hoover (2003; 2009) responds to this problem by distinguishing stationary and non-stationary time series that provide values to *X* and *Y*, and by arguing that non-stationary time series are not subject to the common cause principle unless they are co-integrated. Time series are non-stationary if they grow over time and do not have a fixed (or "stationary") mean. And they are co-integrated if each of them is *I*(1), i.e. integrated of order 1, and if there is a linear combination of them that is *I*(0), i.e. integrated of order 0, where time series or linear combinations of them are *I*(*d*), i.e. integrated of order *d*, if they must be differentiated *d* times to be made stationary.

Hoover's response is convincing to the extent that it explains why neither the common effect principle nor the common cause principle obtains in cases in which $\rho_{XY} \neq 0$ denotes a nonsense correlation: nonsense correlations are correlations between non-stationary time series that are not co-integrated.[10] It is worth mentioning, however, that testing for co-integration is not always easy. Johansen (1988) has developed an empirical procedure that can be applied to test for co-integration, but Cheung and Lai (1993) point to several finite-sample shortcomings of that procedure; and Pagan (1995) points to difficulties in interpreting

[9] Cf. Henschen (2018, section 5) for an elaboration of this second problem.

[10] For a more thorough and critical discussion of Hoover's response, cf. Reiss (2015, chap. 8).

co-integration relationships that stem from the fact that Johansen’s procedure involves estimations of reduced form equations.

References:

Bessler, D. A. and S. Lee. (2002). “Money and Prices: U.S. Data 1869-1914 (a Study with Directed Graphs),” *Empirical Economics* 27(3), pp. 427-46.

Cartwright, N. (1979). “Causal Laws and Effective Strateges”. *Nous* 13(4), pp. 419-437.

Cartwright, N. (1999). *The Dappled World*. Cambridge: CUP.

Cheung, Y.-W. and Lai, K. S. (1993). “Finite-Sample Sizes of Johansen’s Likelihood Ratio Tests for Cointegration.” *Oxford Bulletin of Economics and Statistics* 55, pp. 313-28.

Demiralp, S. and K. D. Hoover. (2003) “Searching for the Causal Structure of a Vector Autoregression.” *Oxford Bulletin of Economics and Statistics* 65, pp. 745-767.

Granger, C. W. J. (1969). “Investigating Causal Relations By Econometric Models and Cross-Spectrum Methods.” *Econometrica* 37(3), pp. 424-438.

Granger, C. W. J. (1980). “Testing for Causality: A Personal Viewpoint.” *Journal of Economic Dynamics and Control* 2(4), pp. 329-352.

Hendry, D. (1980). “Econometrics – Alchemy or Science?” *Economica* 47(188), pp. 387-406.

Henschen, T. (2018). “The in-principle inconclusiveness of causal evidence in macroeconomics.” *European Journal for Philosophy of Science* 8, pp. 709-733.

Hoover, K. D. (1993). “Causality and Temporal Order in Macroeconomics or Why Even Economists Don't Know How to Get Causes from Probabilities.” *The British Journal for Philosophy of Science* 44(4), pp. 693-710.

Hoover, K. D. (2001). *Causality in Macroeconomics*, Cambridge: CUP.

Hoover, D. (2003). “Nonstationary Time Series, Cointegration, and the Principle of Common Cause.” *The British Journal for Philosophy of Science* 54, 527-551.

Hoover, K. D. (2009). “Probability and Structure in Econometric Models.” In C. Glymour et al (eds.), *Logic, Methodology, and Philosophy of Science*. London: College Publications, pp. 497-513.

Jacobs, R. L., Leamer, E. E., Ward, M. P. (1979). “Difficulties with Testing for Causation.” *Economic Inquiry* 17, pp. 401-413.

Johansen, S. (1988). “Statistical Analysis of Cointegration Vectors.” *Journal of Economic Dynamics and Control* 12, pp. 231-254.

Maddala, G. S. and K. Lahiri (^{4}2009). *Introduction to Econometrics*. Chichester: Wiley & Sons.

Mayo, D.G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Norton, J. D. (2011). "Challenges to Bayesian Confirmation Theory." In: P. S. Bandyopadhyay and M. R. Forster (eds.), *Handbook of the Philosophy of Science. Vol. 7: Philosophy of Statistics*. Amsterdam: Elsevier.

Pagan, A. (1995). "Three Methodologies: An Update." In L. Oxley et al. (eds.), *Surveys in Econometrics*. Oxford: Basil Blackwell, pp. 30-41.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, MA: Cambridge University Press.

Reiss, J. (2015). *Causation, Evidence, and Inference*. London: Routledge.

Sargent, T. J. (1977). "Response to Gordon and Ando." In C. A. Sims (ed.), *New Methods in Business Cycle Research*. Minneapolis: Federal Reserve Bank of Minneapolis.

Sims, C. A. (1972). "Money, Income and Causality." *American Economic Review* 62(4), pp. 540-552.

Sims, C. A. (1980a). "Macroeconomics and Reality." *Econometrica* 48, pp. 1-48.

Sims, C. A. (1980b), "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered". *The American Economic Review*, Vol. 70, No. 2, pp. 250-257.

Sober, E. (2001). "Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause." *The British Journal for the Philosophy of Science* 52, pp. 331-346.

Spanos, A. (2000). "Revisiting data mining: 'hunting' with or without a license." *Journal of Economic Methodology* 7: 2, 231-264.

Spirtes, P., Glymour, C., Scheines, R. (1993). *Causation, Prediction and Search*. New York: Springer.

Spohn, W. (1980). "Stochastic Independence, Causal Independence, and Shieldability." *Journal of Philosophical Logic* 9, pp. 73-99.

Spohn, W. (1983). "Probabilistic causality: from Hume via Suppes to Granger." In M. C. Galavotti and G. Gambetta (eds.), *Causalità e modelli probabilistici*. Bologna: Cooperativa Libraria Universitaria.

Spohn, W. (2012). *The Laws of Belief*. Oxford: OUP.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland.

Woodward, J. (2003). *Making Things Happen: a Causal Theory of Explanation*. Oxford: OUP.

Zellner, A. (1979). “Causality and econometrics.” *Carnegie-Rochester Conference Series on Public Policy*. Elsevier, vol. 10(1), pp. 9-54.

Zellner, A. (1988). “Causality and causal laws in economics.” *Journal of Econometrics* 39, pp. 7-21.