Statistics and Computing (1992) **2**, 91–95

A statistical semantics for causation

JUDEA PEARL and THOMAS S. VERMA

Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, CA 90024, USA

Received January 1991 and accepted September 1991

We propose a model-theoretic definition of causation, and show that, contrary to common folklore, genuine causal influences can be distinguished from spurious covariations following standard norms of inductive reasoning. We also establish a sound characterization of the conditions under which such a distinction is possible. Finally, we provide a proof-theoretical procedure for inductive causation and show that, for a large class of data and structures, effective algorithms exist that uncover the direction of causal influences as defined above.

Keywords: Causality, induction, learning

1. The model

We view the task of causal modeling as an identification game which scientists play against nature. Nature possesses stable causal mechanisms which, on a microscopic level, are deterministic functional relationships between variables, some of which are unobservable. These mechanisms are organized in the form of an acyclic schema which the scientist attempts to identify.

Definition 1. A *causal model* over a set of variables U is a directed acyclic graph (DAG) D, the nodes of which denote variables, and the links denote direct binary causal influences.

The causal model serves as a blueprint for forming a 'causal theory'—a precise specification of how each variable is influenced by its parents in the DAG. Here we assume that nature is at liberty to impose arbitrary functional relationships between each effect and its causes and then to weaken these relationships by introducing arbitrary (yet mutually independent) disturbances. These disturbances reflect 'hidden' or unmeasurable conditions and exceptions which nature chooses to govern by some undisclosed probability function.

Definition 2. A *causal theory* is a pair $T = \langle D, \Theta_D \rangle$ containing a causal model D and a set of parameters Θ_D compatible with D. Θ_D assigns a function $x_i = f_i[pa(x_i), \epsilon_i]$ and a probability measure g_i, to each $x_i \in U$, where $pa(x_i)$ are the parents of x_i in D and each ϵ_i is a random disturbance distributed according to g_i, independently of the other ϵs and of $\{x_j\}_{j=1}^{i-1}$.

The requirement of independence renders the disturbances 'local' to each family; disturbances that influence several families simultaneously will be treated explicitly as 'latent' variables (see Definition 3 below).

Once a causal theory T is formed, it defines a joint probability distribution $P(T)$ over the variables in the system, and this distribution reflects some features of the causal model (e.g., each variable must be independent of its grandparents, given the values of its parents). Nature then permits the scientist to inspect a select subset O of 'observed' variables, and to ask questions about the probability distribution over the observables, but hides the underlying causal theory as well as the structure of the causal model. We investigate the feasibility of recovering the topology of the DAG from features of the probability distribution.[1]

[1]This formulation employs several idealizations of the actual task of scientific discovery. It assumes, for example, that the scientist obtains the distribution directly, rather than events sampled from the distribution. This assumption is justified when a large sample is available, sufficient to reveal all the dependencies embedded in the distribution. Additionally, we assume that the observed variables actually appear in the original causal theory and are not some aggregate thereof. Aggregation might result in feedback loops, which we do not discuss in this paper. Our theory also takes variables as the primitive entities in the language, not events which permits us to include 'enabling' and 'preventing' relationships as part of the mechanism.

2. Model preferences (Occam's razor)

In principle, with no restriction on the type of models considered, the scientist is unable to make any meaningful assertions about the structure of the underlying model. For example, he/she can never rule out the possibility that the underlying model is a complete (acyclic) graph; a structure that, with the right choice of parameters can *mimic* (see Definition 4 below) the behavior of any other model, regardless of the variable ordering. However, following the standard method of scientific induction, it is reasonable to rule out any model for which we find a simpler, *less expressive* model, equally consistent with the data (see Definition 6 below). Models that survive this selection are called *minimal* models and, with this notion, we construct our definition of *inductive causation*: 'A variable X is said to have a direct causal influence on a variable Y if a unidirected edge exists in all minimal models consistent with the data'.

Definition 3. A *latent structure* is a pair $L = \langle D, O \rangle$ containing a causal model D over U and a set $O \subseteq U$ of observable variables.

Definition 4. $L = \langle D, O \rangle$ is *preferred* to $L' = \langle D', O \rangle$, written $L \preceq L'$, if and only if D' can *mimic* D over O, i.e. for every Θ_D there exists a $\Theta'_{D'}$ such that $P_{[O]}(\langle D', \Theta'_{D'} \rangle) = P_{[O]}(\langle D, \Theta_D \rangle)$. Two latent structures are *equivalent*, written $L' \equiv L$, if and only if $L \preceq L'$ and $L \succeq L'$.

Definition 5. A latent structure L is *minimal* with respect to a class $\mathscr{L}$ of latent structures if and only if for every $L' \in \mathscr{L}$, $L \equiv L'$ whenever $L' \preceq L$.

Definition 6. $L = \langle D, O \rangle$ is *consistent* with a sampled distribution $\hat{P}$ over O if D can accommodate some theory that generates $\hat{P}$, i.e. there exists a Θ_D such that $P_{[O]}(\langle D, \Theta_D \rangle) = \hat{P}$.

Definition 7 (Induced Causation). Given $\hat{P}$, a variable C has a *direct causal influence* on E if and only if a path from C to E exists in every minimal latent structure consistent with $\hat{P}$.

We view this definition as normative, because it is based on one of the least disputed norms of scientific investigation: Occam's razor in its semantical casting. However, as with any scientific inquiry, we make no claims that this definition is guaranteed always to identify stable physical mechanisms in nature; it identifies the only mechanisms we can plausibly induce from non-experimental data.

[2]It is possible to show that, if the parameters are chosen at random from any reasonable distribution, then any unstable distribution has measure zero (Spirtes, Glymour and Scheines, 1989). Stability precludes deterministic constraints as well as aggregated variables.

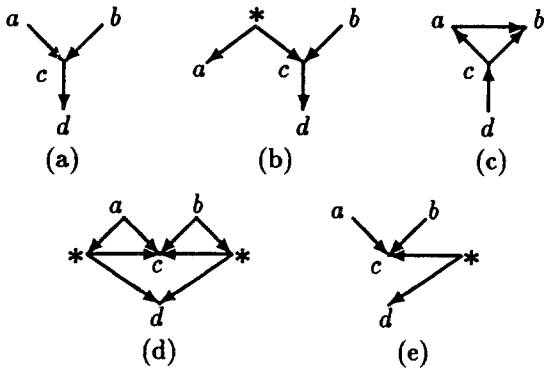

Fig. 1. *Causal models illustrating the soundness of $c \rightarrow d$. The node (*) represents a hidden variable*

As an example of a causal relation that is identified by the definition above, imagine that observations taken over four variables $\{a, b, c, d\}$ reveal only two vanishing dependencies: 'a is independent of b' and 'd is independent of $\{a, b\}$ given c' (plus those that logically follow from the two). This dependence pattern would be typical, for example, of the following variables: *a = having cold, b = having hay-fever, c = having to sneeze, d = having to wipe one's nose.* It is not hard to show that any model which explains the dependence between c and d by an arrow from d to c, or by a hidden common cause between the two, cannot be minimal, because any such model would be able to outmimic the one shown in Fig. 1(a). We conclude, therefore, that the observed dependencies imply a direct causal influence from c to d. Some minimal models (b) and non-minimal models (c and d) consistent with the observations are shown. However, model (e) is inconsistent because it cannot account for the observed marginal dependence between b and d.

3. Proof theory

It turns out that while the minimality principle is sufficient for forming a normative and operational theory of causation, it does not guarantee that the search through the vast space of minimal models would be computationally practical. If nature truly conspires to conceal the structure of the underlying model she could annotate the model with a distribution that matches many minimal models, having totally disparate structures. To facilitate an effective proof theory, we rule out such eventualities, and impose a restriction on the distribution called *stability*. It conveys the assumption that all vanishing dependencies are structural, not formed by incidental equalities of numerical parameters.[2]

Definition 8. Let $I(P)$ denote the set of all conditional independence relationships embodied in P. A probability distribution $\hat{P}$ is *stable* if there exists a dag D such that $\hat{P}$

precisely embodies the independencies dictated by D, i.e. there exists a set of parameters Θ_D such that for any other set Θ'_D we have: $I(P(\langle D, \Theta_D \rangle)) \subseteq I(\hat{P}) \subseteq I(P(\langle D, \Theta'_D \rangle))$.

With the added assumption of stability, every distribution has a unique causal model (up to equivalence), as long as there are no hidden variables (Verma and Pearl, 1990). The search for the minimal model then boils down to recovering the structure of the underlying DAG from probabilistic dependencies that perfectly reflect this structure (see Pearl and Verma, 1987 and Pearl, 1988, for a characterization of these dependencies). This search is exponential in general, but simplifies significantly when the underlying structure is sparse (for such algorithms, see Verma and Pearl, 1990; Spirtes and Glymour, 1991).

4. Recovering latent structures

When nature decides to 'hide' some variables, the observed distribution $\hat{P}$ need no longer be stable relative to the observable set O, i.e. $\hat{P}$ may result from many equivalent minimal latent structures, each containing any number of hidden variables. Fortunately, rather than having to search through this unbounded space of latent structures, it turns out that for every latent structure L, there is an equivalent latent structure called the *projection* of L on O in which every unobserved node is a root node with exactly two observed children.

Definition 9. A latent structure $L_{[O]} = \langle D_{[O]}, O \rangle$ is a *projection* of another latent structure L if and only if

(1) every unobservable variable of $D_{[O]}$ is a parentless common cause of exactly two non-adjacent observable variables;

(2) for every stable distribution P generated by L, there exists a stable distribution P' generated by $L_{[O]}$ such that $I(P_{[O]}) = I(P'_{[O]})$.

Theorem 1. Any latent structure has at least one projection (identifiable in linear time).

(Proofs can be found in Verma, 1992.)

It is convenient to represent projections by bidirectional graphs with only the observed variables as vertices (i.e. leaving the hidden variables implicit). Each bidirected link in such a graph represents a common hidden cause of the variables corresponding to the link's end-points.

Theorem 1 renders our definition of induced causation (Definition 7) operational; we will show (Theorem 2 below) that if a certain link[4] exists in a distinguished projection of any minimal model of $\hat{P}$, it must indicate the existence of a causal path in every minimal model of $\hat{P}$. Thus the search reduces to finding a projection of any minimal model of $\hat{P}$ and identifying the appropriate links. Remarkably, these links can be identified by a simple procedure, the IC algorithm which is no more complex than that which recovers the unique minimal model in the case of fully observable structures.

IC Algorithm (Inductive Causation).
Input: $\hat{P}$, a sampled distribution.
Output: $core(\hat{P})$, a marked hybrid acyclic graph.

(1) For each pair of variables a and b, search for a set S_{ab} such that (a, S_{ab}, b) is in $I(\hat{P})$, namely a and b are independent in $\hat{P}$, conditioned on S_{ab}. If there is no such S_{ab}, place an undirected link between the variables.

(2) For each pair of non-adjacent variables a and b with a common neighbor c, check if $c \in S_{ab}$.
If it is, then continue.
If it is not, then add arrowheads pointing at c, (i.e. $a \rightarrow c \leftarrow b$).

(3) Form $core(\hat{P})$ by recursively adding arrowheads according to the following two rules:[3]
If $\overline{ab}$ and there is a strictly directed path from a to b then add an arrowhead at b.
If a and b are not adjacent but $\vec{ac}$ and $c - b$, then direct the link $c \rightarrow b$.

(4) If $\vec{ab}$ then mark every unidirected link $b \rightarrow c$ in which c is not adjacent to a.

The result of the IC algorithm is a substructure called $core(\hat{P})$ in which every marked unidirected arrow $X \rightarrow Y$ stands for the statement: 'X is a direct cause of Y (in all minimal latent structures consistent with the data)'. We call these relationships *genuine* causes (e.g., $c \rightarrow d$ in Fig. 1(a)).

Theorem 2. If every link of the directed path $C \rightarrow^* E$ is marked in $core(\hat{P})$ then C has a causal influence on E according to $\hat{P}$.

Theorem 3. If $core(\hat{P})$ contains a bidirectional link $E_1 \leftrightarrow E_2$, then there is a common cause X influencing both E_1 and E_2, and no direct causal influence between the two, in every minimal latent structure consistent with $\hat{P}$.

5. Summary and intuition

For the sake of completeness we now present explicit definitions of potential and genuine causation, as they emerge from Theorem 2 and the IC algorithm. Additional conditions, sufficient for the determination of spurious and genuine causes, with and without temporal information, can be found in Pearl and Verma (1991).

[3] $\overline{ab}$ denotes adjacency, $\vec{ab}$ denotes either $a \rightarrow b$ or $a \leftrightarrow b$.
[4] In a hybrid graph links may be undirected, unidirected or bidirected.

Definition 11 (Potential Cause). A variable X has a *potential causal influence* on another variable Y (*inferable* from $\hat{P}$), if

(1) X and Y are dependent in every context.
(2) There exists a variable Z and a context S such that
  (i) X and Z are independent given S
  (ii) Z and Y are dependent given S

Definition 12 (Genuine Cause). A variable X has a *genuine causal influence* on another variable Y if there exists a variable Z such that either:

1. X is a potential cause of Y and there exists a context S satisfying:
  (i) Z is a potential cause of X.
  (ii) Z and Y are dependent given S.
  (iii) Z and Y are independent given $S \cup X$.

2. X and Y are in the transitive closure of the relation defined by Part 1, that is, there exist $k \geq 1$ variables, $W_1, \ldots, W_k$ such that X has a genuine causal influence on W_1 and W_i has a genuine causal influence on W_{i+1} for all $k > i \geq 1$ and W_k has a genuine causal influence on Y, all defined by Part 1.

Definition 11 was formulated in Pearl (1990) as a relation between events (rather than variables) with the added condition $P(Y \mid X) > P(Y)$ in the spirit of Suppes (1970). Condition (1) in Definition 12 may be established either by statistical methods (per Definition 11) or by other sources of information, e.g., experimental studies or temporal succession (i.e. that Z precedes X in time). When temporal information is available, as it is assumed in the formulations of Suppes (1970), Granger (1988) and Spohn (1983), then every link constructed in step 1 of the IC algorithm corresponds to a potential cause (genuine or *spurious cause* in Suppes terminology). In such cases, Definition 12 can be used to distinguish genuine from spurious causes without the usual requirement that all causally relevant background factors be measurable.

The intuition behind our definitions (and the IC recovery procedure) is rooted in Reichenbach's (1956) 'common cause' principle stating that if two events are correlated, but one does not cause the other, then there must be causal explanation to both of them, an explanation that renders them conditionally independent. As it turns out, the pattern that provides us with information about causal directionality is not the 'common cause' but rather the 'common effect'. The argument goes as follows: If we create conditions (fixing S_{ab}) where two variables, a and b, are each correlated with a third variable c but are independent of each other, then the third variable cannot act as a cause of a or b; it must be either their common effect, $a \rightarrow c \leftarrow b$, or be associated with a and b via common causes, forming a pattern such as $a \leftrightarrow c \leftrightarrow b$. This is indeed the eventuality that permits our algorithm to begin orienting edges in the graph (step 2), and assign arrowheads pointing at c. Another explanation of this principle appeals to the perception of 'voluntary control' (Pearl, 1988, p. 396). The reason why people insist that the rain causes the grass to become wet, and not the other way around, is that they can find other means of getting the grass wet, totally independent of the rain. Transferred to our chain $a-c-b$, we can preclude c from being a cause of a if we find another means of potentially controlling c without affecting a, namely b.

The notion of genuine causation also rests on the 'common effect' principle: Two causal events do not become dependent simply by virtue of predicting a common effect. Thus, a series of spurious associations, each resulting from a separate common cause, is not transitive; it predicts independence between the first and last variables in the chain. For example, if I hear my sprinklers turn on, it suggests that my grass is wet, but not that the parking lot at the local supermarket is wet even though the latter two events are highly correlated by virtue of a common cause in the form of rain.[5] Therefore, if correlation is measured between my sprinkler and the wetness of the parking lot then there ought to be a non-spurious causal connection between the wetness of my grass and that of the parking lot (such as the water saturating my lawn, running off into the gutter and into the parking lot).

6. Conclusions

The results presented in this paper dispel the claim that statistical analysis can never distinguish genuine causation from spurious covariation (Otte, 1981; Cliff, 1983; Holland, 1986; Gardenfors, 1988; Cartwright, 1989). We show that certain patterns of dependencies dictate a causal relationship between variables, one that cannot be attributed to hidden causes lest we violate one of the basic maxims of scientific methodology: the semantical version of Occam's razor.

On the practical side, we have shown that the assumptions of model minimality and 'stability' (no accidental independencies) lead to an effective algorithm for recovering causal structures, transparent as well as latent. Simulation studies conducted at our laboratory show that networks containing 20 variables require less than 5000 samples to have their structure recovered by the algorithm. Another result of practical importance is the following: Given a proposed causal theory of some phenomenon, our algorithm can identify those causal relationships (or the lack thereof) that could potentially be substantiated by observational studies, and those whose directionality might require determination by controlled, manipulative experiments.

[5]Apparently this lack of transitivity has not been utilized by path analysts.

From the methodological viewpoint, our results should settle some of the ongoing disputes between the descriptive and structural approaches to theory formation (Freedman, 1987). It shows that the methodology governing path-analytic techniques is legitimate, faithfully adhering to the traditional norms of scientific investigation. At the same time, our results also explicate the assumptions upon which these techniques are based, and the conditions that must be fulfilled before claims made by these techniques can be accepted.

Acknowledgements

This work was supported, in part, by NSF grant IRI-88-2144, and NRL grant N000-89-J-2007. T. S. Verma was supported by an IBM graduate fellowship.

References

Cartwright, N. (1989) *Nature Capacities and Their Measurements*. Clarendon Press, Oxford.

Cliff, N. (1983) Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, **18**, 115–126.

Freedman (1987) As others see us: a case study in path analysis (with discussion). *Journal of Educational Statistics*, **12**, 101–223.

Gardenfors, P. (1988) Causation and the dynamics of belief, in *Causation in Decision, Belief Change and Statistics II*, Harper, W. L. and Skyrms, B. (eds), Kluwer Academic Publishers, Dordrecth, pp. 85–104.

Glymour, C., Scheines, R., Spirtes, P. and Kelly, K. (1987) *Discovering Causal Structure*, Academic Press, New York.

Granger, C. W. J. (1988) Causality testing in a decision science, in *Causation in Decision, Belief Change and Statistics I*, Harper, W. L. and Skyrms, B. (eds), Kluwer Academic Publishers, Dordrecht, pp. 1–20.

Holland, P. (1986) Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945–960.

Kautz, H. (1987) A formal theory of plan recognition. PhD thesis, University of Rochester, Rochester, NY.

Otte, R. (1981) A critique of Suppes' theory of probabilistic causality. *Synthese*, **48**, 167–189.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1990) Probabilistic and qualitative abduction, in *Proceedings of AAAI Spring Symposium on Abduction*, Stanford, March 27–29, pp. 155–158.

Pearl, J. and Verma, T. S. (1987) The logic of representing dependencies by directed acyclic graphs. *Proceedings of AAAI-87*, Seattle, Washington, pp. 347–379.

Pearl, J. and Verma, T. S. (1991) A theory of inferred causation. In Allen, J. A., Fikes, R. and Sandwall, E. (eds), *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pp. 441–452. Morgan Kaufmann, San Mateo.

Reichenbach, H. (1956) *The Direction of Time*, University of California Press, Berkeley.

Simon, H. (1954) Spurious correlations: a causal intepretation. *Journal American Statistical Association*, **49**, 469–492.

Spirtes, P. and Glymour, C. (1991) An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, **9**, 62–72.

Spirtes, P., Glymour, C. and Scheines, R. (1989) Causality from probability. Technical Report CMU-LCL-89-4, Department of Philosophy, Carnegie-Mellon University.

Spohn, W. (1983) Deterministic and probabilistic reasons and causes. *Erkenntnis*, **19**, 371–396.

Suppes, P. (1970) *A Probabilistic Theory of Causation*, North-Holland, Amsterdam.

Verma, T. S. and Pearl, J. (1990) Equivalence and synthesis of causal models, in *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, Cambridge, MA, pp. 220–227. Also published, North-Holland, Amsterdam (1991) 255–268.

Verma, T. S. (1992) Invariant properties of causal models. *Technical Report* R-134, UCLA Cognitive Systems Laboratory.