Computer Science Department Technical Report
University of California
Los Angeles, CA 90024-1596

A FORMAL THEORY OF INDUCTIVE CAUSATION

Judea Pearl
Thomas Verma

July 1991
CSD-910024

given set of observations, and the meaning of explanation is intimately related to the notion of causation.

Most AI works have given the term "cause" a procedural semantics, attempting to match the way people use it in reasoning tasks, but were not concerned with the experience that prompts people to believe that "a causes b", as opposed to, say, "b causes a" or "c causes both a and b." The question of choosing an appropriate causal ordering received most attention in qualitative physics, where some connectives attain directionality despite the symmetrical nature of the functional equations [Kuipers 84] [Forbus & Gentner 86]. In some systems causal ordering is defined as the ordering at which subsets of variables can be solved independently of others [Iwasaki & Simon 86] or the way a disturbance is propagated from one variable to others [de Kleer & Brown 86]. Yet these choices are made as a matter of convenience, to fit the structure of a given theory, and do not reflect features of the empirical environment which compelled the formation of the theory.

An empirical semantics for causation is important for several reasons. First, by tracing empirical origins we gain a better understanding of the meaning of utterances such as "a explains b", "a suggests b", "a tends to cause b", and "a actually caused b" , etc. Second, an intelligent system attempting to build a workable model of its environment cannot rely exclusively on preprogrammed causal knowledge, but must be able to translate direct observations to cause-and-effect relationships.

Temporal precedence is normally assumed essential for defining causation. Suppes [Suppes 70], for example, introduced a probabilistic definition of causation with an explicit requirement that a cause precedes its effect in time. [Shoham 88] makes an identical assumption. In this paper we propose a non-temporal semantics, one that determines the directionality of causal influences without resorting to temporal information, in the spirit of [Simon 54] and [Glymour et al. 87]. Such semantics should be applicable, therefore, to the organization of concurrent events or events whose chronological precedence cannot be determined with precision, (e.g. "old age explains disabilities"). Motivated by the observation that most causal utterances describe probabilistic relations (e.g. "reckless driving causes accidents") rather than functional or deterministic dependencies, we base our formalization on probability theory. However, given that statistical analysis is driven by covariation, not causation, we must still identify the asymmetries that prompt people to perceive causal structures in statistical data, and we must find a computational model for such perception. Our results will be discussed in Section 2, while the formal treatment is left for the Appendix.

TECHNICAL REPORT
R-155 (I)
October, 1990
Submitted.

A Formal Theory of Inductive Causation*

Key words: causality, induction, learning

Judea Pearl
< *judea@cs.ucla.edu* >

TS Verma†
< *verma@cs.ucla.edu* >

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

Abstract

This paper concerns the empirical basis of causation, and addresses the following issues:

1. the asymmetries that might prompt people to perceive causal directionality in uncontrolled observations,
2. the task of inferring causal models from these asymmetries, and
3. whether the models inferred tell us anything useful about the causal mechanisms that underly the observations.

We propose a model-theoretic extension of the Reichenbach-Suppes 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 complete 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.

1 Introduction

The study of causation is central to the understanding of human reasoning. Tasks involving changing environments require causal theories which make formal distinctions between causation and logical implication [Shoham 88] [Lifschitz 87]. In applications such as diagnosis [Patil & Schwartz 82] [Reiter 87], qualitative physics [Bobrow 85], and plan recognition [Wilensky 83] [Kautz 87], a central task is that of finding a satisfactory *explanation* to a

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

2 Summary of Major Results

2.1 The Model

We pretend that Nature possesses "true" cause-and-effect relationships and that these relationships are organized in the form of a a directed acyclic graph (dag); each node represents a variable in the domain, and the parents of that node correspond to its direct causes, as designated by Nature. This dag (called a "causal model") serves as a blue print 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. The requirement of independence renders these disturbances "local" to each family; disturbances that influence several families simultaneously will be treated explicitly as "latent" variables.

Once a causal theory is formed, it defines a joint probability distribution over the variables in the system, and this distribution reflects some features of the causal model (e.g., each variable is 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]

2.2 Model preferences (Occam's razor)

In principle, with no restriction whatsoever on the type of models considered, the scientist is unable to make any categorical assertion about the structure of the underlaying 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 the behavior

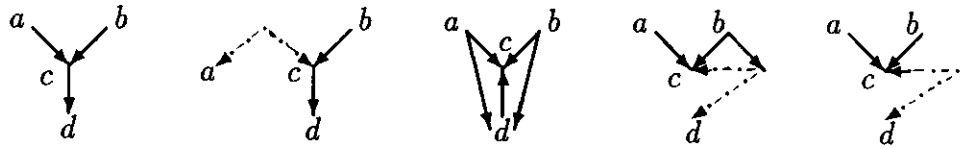
[1] This formulation employs several idealizations of the actual task of scientific discovery. It assumes, for example, that the scientist obtains the distribution directly, rather than events sampled from the distribution. This assumption is justified when a large sample is available, sufficient to 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.

of every causal theory. However, following the standard method of scientific induction, it is reasonable to rule out every model for which we find a simpler, *less expressive model*, equally consistent with the data. 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 uni-directed edge exists in all minimal models consistent with the data"

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 actually identifies stable physical mechanisms; it identifies the only mechanisms inducible from non-manipulative data.

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 ones 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. We conclude therefore that the observed dependencies imply a direct causal influence from c to d. Some minimal (1 and 2) and non-minimal (3 and 4) models consistent with the observations are shown below. However, (5) is inconsistent, because it cannot account for the observed marginal dependence between b and d.



2.3 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 underlying model with a distribution that matches many minimal models, having totally distinct 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] With the added assumption of stability, every distribution has a unique causal model (up to equivalence), as long as there are no hidden variables. 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. This search is exponential in general, but simplifies significantly when the underlying structure is sparse (see [Spirtes, Glymour & Scheines 89] and [Verma & Pearl 90b] for such algorithms).

2.4 Recovering Latent Structures

When Nature decides to "hide" some variables, the observed distribution $\hat{P}$ may have many minimal models (called "latent structures"), each containing any number of hidden variables. Fortunately, rather then having to search through the vast space of latent structures, it turns out that every such structure can be encoded parsimoniously using a bi-directed graph with only the observed variables as vertices. Each bi-directed link in such a graph represents a common hidden cause of the variables corresponding to the link's end points. If we now identify the arrowheads that are shared by all these bi-directed graphs, we obtain the sum total of all causal statements that can be "proven" from the observed distribution.

Remarkably, these arrowheads can be identified by a simple procedure, the IC-algorithm, that is not more complex than searching for the unique minimal model in the case of fully observable structures. The result of this procedure is a substructure called $\mathbf{core}(\hat{P})$ in which every marked uni-directed 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 previous figure).

The intuition behind our 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

[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 & Scheines 89]. Stability precludes deterministic constraints.

b, are each correlated with a third variable c but are independent of each other, then the third variable cannot be 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 the pattern $a \leftrightarrow c \leftrightarrow b$. 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 88, page 396]. The reason 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 (b) of potentially controlling c without affecting a.

3 Conclusions

The results presented in this paper dispel the claim that statistical analysis can never distinguish genuine causation from spurious covariation [Otte 81]. We show that certain patterns of dependencies dictate a direct 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 assumption of model minimality, together with that of "stability" (no accidental independencies) lead to an effective algorithm of recovering causal structures, transparent as well as latent. Simulation studies conducted at our laboratory show that networks containing twenty 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 in linear time those causal relationships that could potentially be substantiated by observational studies, and those whose directionality can only be determined by controlled, manipulative experiments.

From a methodological viewpoint, our results should settle some of the on going disputes between the descriptive and structural approaches to theory formation [Freedman 87]. 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.

References

[Bobrow 85] *Qualitative Reasoning about Physical Systems*, MIT Press, Cambridge, MA. 1985.

[de Kleer & Brown 86] de Kleer, J.; and Brown, J. S. Theories of causal ordering. *Artificial Intelligence*, 1986, 29(1):33–62.

[Freedman 87] As Others See Us: A Case Study in Path Analysis (with discussion). *Journal of Educational Statistics*, 1987, 12:101–223.

[Forbus & Gentner 86] Forbus, K. D. and Gentner, D., Causal Reasoning about Quantities. *Proceedings Cognitive Science Society*, Amherst, 1986, 196–207.

[Glymour et al. 87] Glymour, C.; Scheines, R.; Spirtes, P.; and Kelly, K. *Discovering Causal Structure*, Academic Press, New York, 1987.

[Iwasaki & Simon 86] Iwasaki, Y.; and Simon H. A. Causality in Device Behavior. *Artificial Intelligence*, 1986, 29(1):3–32.

[Kautz 87] *A formal Theory of Plan Recognition.* PhD thesis, University of Rochester, Rochester, N.Y., May 1987.

[Kuipers 84] Kuipers, B., Commonsense Reasoning about Causality: Deriving Behavior from Structure. *Artificial Intelligence*, 1984, 24(1–3):169–203.

[Lifschitz 87] Formal Theories of Action. *Proceedings 1987 Workshop of the Frame Problem in AI*, Kansas, 1987, 35–57.

[Otte 81] Otte, R. A critique of Suppes' theory of Probabilistic causality. *Synthese* 48:167–189.

[Patil & Schwartz 82] Patil R.S., Szolovitz,P. and Schwartz, W.B., Causal understanding of patient illness in patient diagnosis. *Proceedings of AAAI-82*, 1982.

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

[Pearl 88] Pearl, J. *Probabilistic Reasoning in Intelligent Systems.* Morgan-Kaufman, San Mateo, CA, 1988.

[Reiter 87] Reiter, R., A theory of diagnosis from first principles. *Artificial Intelligence*, 1987, 32(1):57-95.

[Shoham 88] Shoham, Y. *Reasoning About Change.* MIT Press, Boston, MA, 1988.

[Simon 54] Simon, H. Spurious correlations: A causal interpretation. *Journal American Statistical Association*, 1954, 49:469–492.

[Spirtes, Glymour & Scheines 89] Spirtes, P.; Glymour, C.; and Scheines, R. Causality from probability. *Technical Report CMU-LCL-89-4*, Department of Philosophy Carnegie-Mellon University, 1989.

[Suppes 70] Suppes, P. *A Probabilistic Theory of Causation.* North Holland, Amsterdam, 1970.

[Verma & Pearl 90a] Verma, T. S.; and Pearl J. Causal networks: Semantics and expressiveness. *Uncertainty in Artificial Intelligence 4*, Shachter R. D., et al. (Eds). North-Holland, 1990, 69–76.

[Verma & Pearl 90b] Verma, T. S.; and Pearl J. Equivalence and Synthesis of Causal Models. *Proceedings 6th Conference on Uncertainty in AI*, Mass, July 1990, 220-227.

[Wilensky 83] Wilensky, R. *Planning and understanding*, Addison Wesley, 1983.

A Formal Treatment, Definitions and Theorems

A.1 Model Theory

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.*

Definition 2 *A* **latent structure** *is a pair $L = <D, O>$ containing a causal model D over U and a set $O \subseteq U$ of observable variables.*

Definition 3 *A* **causal theory** *is a pair $T = <D, \Theta_D>$ containing a causal model D and a set of parameters Θ_D compatible with D. Θ_D assigns a function $x_i = f_i[\mathbf{pa}(x_i), \epsilon_i]$ and a probability measure g_i, to each $x_i \in U$, where $\mathbf{pa}(x_i)$ 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 ϵ's and of $\{x_j\}_{j=1}^{i-1}$*

Notation: For a causal theory $T = <D, \Theta_D>$ and latent structure $L = <D, O>$,
$P(T)$ denotes the probability distribution generated by T.
$I(D)$ denotes the set of triples (X, Z, Y) of subsets of variables s.t. X and Y are d-separated in D given Z (see [Pearl & Verma 87] for details).
$I(P(T))$ denotes the set of triples (X, Z, Y) s.t. X and Y are independent in P given Z.
$I_{[O]}$ denotes the subset of triples of I in which X, Y and Z are taken from O.
$P_{[O]}(T)$ denotes the marginal of $P(T)$ over O.

Theorem 1 [Verma & Pearl 90a] *For any causal theory $T = <D, \Theta_D>$, $I(D) \subseteq I(P(T))$.*

Definition 4 *$L = <D, O>$ is* **preferred** *to $L' = <D', O>$ written, $L \preceq L'$ iff D' can* mimic *D over O, i.e. for every Θ_D there exists a $\Theta'_{D'}$ s.t. $P_{[O]}(<D', \Theta'_{D'}>) = P_{[O]}(<D, \Theta_D>)$. Two latent structures are* **equivalent**, *written $L' \equiv L$, iff $L \preceq L'$ and $L \succeq L'$.*

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

Theorem 2 *Given $L = <D, O>$ and $L' = <D', O>$, $L \preceq L'$ iff $I_{[O]}(D) \supseteq I_{[O]}(D')$.*

Theorem 3 *When $U = O$, two latent structures are equivalent iff their dags have the same links and same set of uncoupled head-to-head nodes*[3].

Definition 6 *$L = <D, O>$ 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 s.t. $P_{[O]}(<D, \Theta_D>) = \hat{P}$*

Theorem 4 *A latent structure $L = <D, O>$ is consistent with $\hat{P}$ iff $I_{[O]}(D) \subseteq I(\hat{P})$.*

Definition 7 (Induced Causation) *Given $\hat{P}$, a variable C has a* **direct causal influence** *on E iff a link $C \rightarrow E$ exists in every minimal latent structure consistent with $\hat{P}$.*

A.2 Proof Theory

Definition 8 *A latent structure $L_{[O]} = <D_{[O]}, O>$ is a* **projection** *of L iff (1) $L_{[O]} \equiv L$ and (2) every unobservable variable of $D_{[O]}$ is a parentless common cause of exactly two non-adjacent observable variables.*

Theorem 5 *Any latent structure has at least one projection (identifiable in linear time).*

Definition 9 *For any latent structure L,* **core**(L) *is defined as the hybrid graph*[4] *satisfying (1) two nodes are adjacent iff they are adjacent or they have a common unobserved cause in every projection of L, and (2) a link between a and b has an arrowhead pointing at b iff $a \rightarrow b$ or a and b have a common unobserved cause in every projection of L.*

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 $I(a, S_{ab}, b)$ holds.
   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:[5]
   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 $\overrightarrow{ac}$ and $c - b$, then direct the link $c \rightarrow b$.

4. Mark any uni-directed link $a \rightarrow b$ if there is some link with an arrowhead directed at a.

Definition 10 *A causal theory $T = <D, \Theta_D>$ generates a* **stable** *distribution iff it contains no extraneous independences, i.e. $I(P(<D, \Theta_D>)) \subseteq I(P(<D, \Theta'_D>))$ for any set of parameters Θ'_D.*

[3] i.e. converging arrows emanating from non-adjacent nodes.

[4] In a hybrid graph links may be undirected, uni-directed or bi-directed.

[5] $\overline{ab}$ denotes adjacency, $\overrightarrow{ab}$ denotes either $a \rightarrow b$ or $a \leftrightarrow b$.

Theorem 6 *For any latent structure* $L = <D, O>$ *and associated theory* $T = <D, \Theta_D>$ *if* $P(T)$ *is stable then* $\mathbf{core}(L) = \mathbf{core}(P_{[O]}(T))$.

Corollary 1 *If* $C \rightarrow E$ *is marked in* $\mathbf{core}(\hat{P})$ *then* C *is a direct cause of* E *according to* $\hat{P}$.