A THEORY OF INFERRED CAUSATION*

Judea Pearl
< *judea@cs.ucla.edu* >
Cognitive Systems Laboratory
Computer Science Department
University of California
Los Angeles, CA 90024

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

We propose a minimal-model semantics 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. We provide an effective algorithm for inferred causation and show that, for a large class of data the algorithm can uncover the direction of causal influences as defined above. Finally, we address the issue of non-temporal causation.

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 [Geffner, 1989, Lifschitz, 1987, Pearl, 1988a, Shoham, 1988]. In applications such as diagnosis [Patil et al., 1982, Reiter, 1987], qualitative physics [Bobrow, 1985], and plan recognition [Kautz, 1987, Wilensky, 1983], a central task is that of finding a satisfactory *explanation* to a 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 some attention in qualitative physics, where certain interactions attain directionality despite the instantaneous and symmetrical nature of the underlying equations, as in "current causing a voltage drop across the resistor" [Forbus and Gentner, 1986]. In some systems causal ordering is defined as the ordering at which subsets of variables can be solved independently of others [Iwasaki and Simon, 1986], in other systems it follows the way a disturbance is propagated from one variable to others [de Kleer and Brown, 1986]. 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, 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. Second, by tracing empirical origins we stand to obtain an independent gauge for deciding which of the many logics proposed for causal reasoning is sound and/or complete, and which provides a proper account of causal utterances such as "a explains b", "a suggests b", "a tends to cause b", and "a actually caused b", etc.

While the notion of causation is often associated with those of necessity and functional dependence, causal expressions often tolerate exceptions, primarily due to missing variables and coarse description. We say, for example, "reckless driving causes accidents" or "you will fail this course because of your laziness". Suppes [Suppes, 1970] has argued convincingly that most causal utterances in ordinary conversation reflect prob-

*This paper is a modified version of one presented at the Second International Conference conference on the Principles of Knowledge Representation and Reasoning, Cambridge, Massachusetts, April 1991.

abilistic, not categorical relations[1]. Thus, probability theory should provide a natural language for capturing causation [Reichenbach, 1956, Good, 1983]. This is especially true when we attempt to infer causation from (noisy) observations – probability calculus remains an unchallenged formalism when it comes to translating statistical data into a system of revisable beliefs.

However, given that statistical analysis is driven by covariation, not causation, and assuming that most human knowledge derives from statistical observations, we must still identify the clues that prompt people to perceive causal relationships in the data, and we must find a computational model that emulates this perception.

Temporal precedence is normally assumed essential for defining causation, and it is undoubtedly one of the most important clues that people use to distinguish causal from other types of associations. Accordingly, most theories of causation invoke an explicit requirement that a cause precedes its effect in time [Good, 1983, Reichenbach, 1956, Shoham, 1988, Suppes, 1970]. Yet temporal information alone cannot distinguish genuine causation from spurious associations caused by unknown factors. In fact the statistical and philosophical literature has adamantly warned analysts that, unless one knows in advance all causally relevant factors, or unless one can carefully manipulate some variables, no genuine causal inferences are possible [Cartwright, 1989, Cliff, 1983, Eells and Sober, 1983, Fisher, 1953, Gardenfors, 1988, Holland, 1986, Skyrms, 1986][2]. Neither condition is realizable in normal learning environments, and the question remains how causal knowledge is ever acquired from experience.

This paper introduces a minimal-model semantics of causation which provides a plausible account for how causal models could be inferred from observations. Using this semantics we show that genuine causal influences can in many cases be distinguished from spurious covariations and, moreover, the direction of causal influences can often be determined without resorting to chronological information. (Although, when available, chronological information can significantly simplify the modeling task.) 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") in the spirit of Glymour [Glymour et al., 1987] and Simon [Simon, 1954].

[1]See [Dechter and Pearl, 1990] for a treatment of causation in the context of categorical data.

[2]Some of the popular quotes are: "No causation without manipulation", [Holland, 1986], "No causes in, no causes out", [Cartwright, 1989] "No computer program can take account of variables that are not in the analysis", [Cliff, 1983].

This paper is organized as follows. In Section 2 we define the notions of causal models and causal theories, and describe the task of causal modeling as an identification game scientists play against Nature. In Section 3 we introduce the minimal-model semantics of causation and exemplify its operability and plausibility on a simple example. Section 4 identifies conditions under which effective algorithms exist that uncover the structure of casual influences as defined above. One such algorithm (called IC) is introduced in Section 5, and is shown to be sound for the class of stable distributions, even when some variables are not observable[3]. Section 6 extracts from the IC-algorithm the essential conditions under which causal influences are identified and proposes these as independent definitions of genuine influences and spurious associations, with and without temporal information. Section 7 provides an intuitive justification for the definitions proposed in Section 6, showing that our theory conforms to the common understanding of causation as a stipulation of stable behavior under external interventions. The definitions are shown to be in line with accepted standards of controlled experimentation, save for requiring the identification of "virtual" experimental conditions within the data itself. In Section 8 we invoke the "virtual control" metaphor to elucidate how causal relationships can still be ascertained in the absence of temporal information. We then offer an explanation for the puzzling, yet universal agreement between the temporal and the statistical aspects of causation.

2 The Causal Modeling Framework

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** *of a set of variables U is a directed acyclic graph* (dag), *in which each node corresponds to a distinct element of U.*

The nodes of the dag correspond to the variables under analysis, while the links denote direct causal influences among the variables. The 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 perturb these relationships by introducing arbitrary (yet mutually independent) disturbances. These disturbances reflect "hidden" or unmeasurable conditions and exceptions

[3]Proofs can be found in [Verma, 1991].

which Nature chooses to govern by some undisclosed probability function.

Definition 2 *A* **causal theory** *is a pair* $T = <D, \Theta_D>$ *consisting of 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 any preceding variable* $x_j : 0 < j < i$

This requirement of independence renders the disturbances "local" to each parents-child family; disturbances that influence several families simultaneously will be treated explicitly as "latent" variables (see Definition 3).

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 $O \subseteq U$ 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, D, from features of the probability distribution.[4]

3 Model preferences (Occam's razor)

In principle, U being unknown, there is an unbounded number of models that would fit a given distribution, each invoking a different set of "hidden" variables and each connecting the observed variables through different causal relationships. Therefore with no restriction on the type of models considered, the scientist is unable to make any meaningful assertions about the structure underlying the phenomena. Likewise, assuming $U = O$ but lacking temporal information, 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) 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). Models that survive this selection are called "minimal models" and with this notion, we can construct our definition of *inferred causation*:

"A variable X is said to have a causal influence on a variable Y if a strictly directed path from X to Y exists in every minimal model consistent with the data"

Definition 3 *Given a set of observable variables* $O \subseteq U$*, a* **latent structure** *is a pair* $L = <D, O>$ *where* D *is a causal model over* U.

Definition 4 *One latent structure* $L = <D, O>$ *is* **preferred** *to another* $L' = <D', O>$ *(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* $L \succeq L'$.

Note that the preference for simplicity imposed by Definition 4 is gauged by the expressive power of a model, not by its syntactic description. For example, one latent structure $L1$ may invoke many more parameters than $L2$ and still be preferred, if $L2$ is capable of accommodating a richer set of probability distributions over the observables. One reason scientists prefer simpler models is that such models are more constrained, thus more falsifiable; they provide the scientist with less opportunities to overfit the data hindsightedly and, therefore attain greater credibility [Pearl, 1978, Popper, 1959].

We also note that the set of dependencies induced by a causal model provides a measure of its expressive power, i.e., its power of mimicing other models. Indeed, $L1$ cannot be preferred to $L2$ if there is even one observable dependency that is induced by $L1$ and not by $L2$. Thus, tests for preference and equivalence can often be reduced to tests of induced dependencies which, in turn, can be determined directly from the topology of the dags, without ever concerning ourselves with the set of parameters. (For example, see Theorem 1 below and [Frydenberg, 1989, Pearl et al., 1989, Verma and Pearl, 1990]).

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* $L' \preceq L$.

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

Clearly, a necessary (and often sufficient) condition for

[4]This formulation invokes 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.

L to be consistent with $\hat{P}$, is that the structure of L can account for all the dependencies embodied in $\hat{P}$.

Definition 7 (Inferred Causation) *Given $\hat{P}$, a variable C has a* **causal influence** *on E iff there exists a directed path $C \rightarrow^* E$ 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 to always identify stable physical mechanisms in nature; it identifies the only mechanisms we can plausibly infer from non-expiremental data.

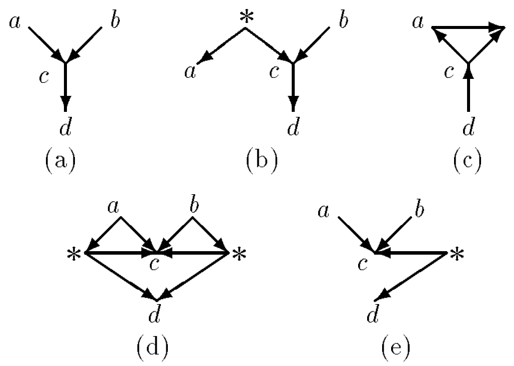

Figure 1: Causal models illustrating the soundness of $c \rightarrow d$. The node $(*)$ represents a hidden variable.

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 two vanishing dependencies: "a is independent of b" and "d is independent of $\{a, b\}$ given c". Assume further that the data reveals *no other* independence, except those that logically follow from these 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 see 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 out-mimic the one shown in Figure 1(a) which reflects all observed independencies. For example, the model of Figure 1(c), unlike that of Figure 1(a), accommodates distributions with arbitrary relations between a and b. Similarly, Figure 1(d) is not minimal as it fails to impose the conditional independence between d and $\{a, b\}$ given c. In contrast, Figure 1(e) is not consistent with the data since it imposes a marginal independence between $\{a, b\}$ and d, which was not observed.

4 Proof Theory and Stable Distributions

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 still annotate that 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" (or "dag-isomorphism" in [Pearl, 1988b]). It conveys the assumption that all vanishing dependencies are structural, not formed by incidental equalities of numerical parameters[5].

Definition 8 *Let I(P) denote the set of all conditional independence relationships embodieded in P. 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.*

With the added assumption of stability, every distribution has a unique causal model (up to equivalence), as long as there are no hidden variables. This uniqueness follows from the fact the structural constraints that an underlying dag imposes upon the probability distribution are equivalent to a finite set of conditional independence relationships asserting that, given its parents, each variable is conditionally independent of all its non-descendents. Therefore two causal models are equivalent (i.e. they can mimic each other) if and only if they relay the same dependency information. The following theorem, which is founded upon the dependency information, states necessary and sufficient conditions for equivalence of causal models which contain no hidden variables.

Theorem 1 [Verma and Pearl, 1990] *When $U = O$, two causal models are equivalent iff their dags have the same links and same set of uncoupled head-to-head nodes*[6].

The search for the minimal model then boils down to recovering the structure of the underlying dag from queries about the dependencies portrayed in that dag. This search is exponential in general, but simplifies significantly when the underlying

[5] 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 et al., 1989]. Stability precludes deterministic constraints. Less restrictive assumptions are treated in [Geiger et al., 1990].

[6] i.e. converging arrows emanating from non-adjacent nodes, such as $a \rightarrow c \leftarrow b$ in Figure 1(a).

structure is sparse (see [Spirtes and Glymour, 1991, Verma and Pearl, 1990] for such algorithms).

Unfortunately, the constraints that a latent structure impose upon the distribution cannot be completely characterized by any set of dependency statements. However, the maximal set of sound constraints can be identified [Verma and Pearl, 1990] and it is this set that permits us to recover sound fragments of latent structures.

5 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 then having to search through this unbounded space of latent structures, it turns out that for every latent structure L, there is a dependency-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 $L_{[O]} = <D_{[O]}, O>$ is a* **projection** *of another latent structure L iff*

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

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

Theorem 2 renders our definition of inferred causation (Definition 7) operational; we will show (Theorem 3) that if a certain link 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, that is not 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: $\mathbf{core}(\hat{P})$ 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 $\mathbf{core}(\hat{P})$ by recursively adding arrowheads according to the following two rules:[7] 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. If a and b are not adjacent but $\overrightarrow{ac}$ and $c \rightarrow b$, then mark the link $c \rightarrow b$.

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 has a causal influence on Y (in all minimal latent structures consistent with the data)". We call these relationships "genuine" causal influences (e.g. $c \rightarrow d$ in previous Figure 1a).

Definition 10 *For any latent structure L,* $\mathbf{core}(L)$ *is defined as the hybrid graph[8] satisfying (1) two nodes are adjacent in* $\mathbf{core}(L)$ *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.*

Theorem 3 *(soundness) For any latent structure $L = <D, O>$ and an associated theory $T = <D, \Theta_D>$ if $P(T)$ is stable then every arrowhead identified by IC is also in* $\mathbf{core}(L)$.

Corollary 1 *If every link of the directed path $C \rightarrow^* E$ is marked in* $\mathbf{core}(\hat{P})$ *then C has a causal influence on E according to $\hat{P}$.*

---

[7] $\overline{ab}$ denotes adjacency, i.e. $a - b$, $a \rightarrow b$, $a \leftarrow b$ or $a \leftrightarrow b$; $\overrightarrow{ab}$ denotes either $a \rightarrow b$ or $a \leftrightarrow b$.

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

6 Probabilistic Definitions for Causal Relations

The IC-algorithm takes a distribution $\hat{P}$ and outputs a dag, some of its links are marked uni-directional (denoting genuine causation), some are unmarked uni-directional (denoting potential causation), some are bi-directional (denoting spurious association) and some are undirected (denoting relationships that remain undetermined). The conditions which give rise to these labelings constitute operational definitions for the various kinds of causal relationships. In this section we present explicit definitions of potential and genuine causation, as they emerge from Theorem 3 and the IC-algorithm. Note that in all these definitions, the criterion for causation between two variables, X and Y, will require that a third variable Z exhibit a specific pattern of interactions with X and Y. This is not surprising, since the very essence of causal claims is to stipulate the behavior of X and Y under the influence of a third variable, one that corresponds to an external control of X. Therefore, our definitions are in line with the paradigm of "no causation without manipulation" [Holland, 1986]). The difference is only that the variable Z, acting as a virtual control of X, must be identified within the data itself. The IC-algorithm provides a systematic way of searching for variables Z that qualify as virtual controls.

Detailed discussions of these definitions in terms of virtual control are given in Sections 7 and 8.

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*

Note that this definition precludes a variable X from being a potential cause of itself or of any other variable which functionally determines X.

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

1. *X and Y are dependent in any context 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 $S \cup X$,*

   or,
2. *X and Y are in the transitive closure of rule 1.*

Definition 13 (Spurious Association) *Two variables X and Y are* **spuriously associated** *if they are dependent in some context S and there exists two other variables Z_1 and Z_2 such that:*

1. *Z_1 and X are dependent given S*
2. *Z_2 and Y are dependent given S*
3. *Z_1 and Y are independent given S*
4. *Z_2 and X are independent given S*

Succinctly, using the predicates I and $\neg I$ to denote independence and dependence respectively, the conditions above can be written:

1. $\neg I(Z_1, X|S)$
2. $\neg I(Z_2, Y|S)$
3. $I(Z_1, Y|S)$
4. $I(Z_2, X|S)$

Definition 11 was formulated in [Pearl, 1990] as a relation between events (rather than variables) with the added condition $P(Y|X) > P(Y)$ in the spirit of [Good, 1983, Reichenbach, 1956, 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 most theories of causality ([Granger, 1988, Spohn, 1983, Suppes, 1970]), then Definitions 12 and 13 simplify considerably because every variable preceding and adjacent to X now qualifies as a "potential cause" of X. Moreover, adjacency (i.e. condition 1 of Definition 11) is not required as long as the context S is confined to be earlier than S. These considerations lead to simpler conditions distinguishing genuine from spurious causes as shown next.

Definition 14 (Genuine Causation with temporal information) *A variable X has a causal influence on Y if there is a third variable Z and a context S, both occurring before X such that:*

1. $\neg I(Z, Y|S)$
2. $I(Z, Y|S \cup X)$

Definition 15 (Spurious Association with temporal information) *Two variables X and Y are* **spuriously associated** *if they are dependent in some context S, X precedes Y and there exists a variable Z satisfying:*

1. $I(Z, Y|S)$
2. $\neg I(Z, X|S)$

7 Causal Intuition and Virtual Experiments

This section explains how the formulation introduced above conforms to common intuition about causation and, in particular, how symmetric probabilistic dependencies can be transformed into judgements about causal influences. We shall first uncover the intuition behind Definition 14, assuming the availability of temporal information, then (in Section 8) generalize to non temporal data, per Definition 12.

The common intuition about causation is captured by the heuristic definition [Rubin, 1989]: "X is a cause for Y if an external agent interfering only with X can affect Y" .

Thus, causal claims are much bolder than those made by probability statements; not only do they summarize relationships that hold in the distribution underlying the data, but they also predict relationships that should hold when the distribution undergoes changes, such as those inferable from external intervention. The claim "X causes Y" asserts the existence of a *stable* dependence between X and Y, one that cannot be attributed to some prior cause common to both, and one that should be preserved when an exogenous control is applied to X.

This intuition requires the formalization of three notions:

1. That the intervening agent be "external" (or "exogenous")
2. That the agent can "affect" Y
3. That the agent interferes "only" with X

If we label the behavior of the intervening agent by a variable Z, then these notions can be given the following probabilistic explications:

1. **Externality of** Z: Variations in Z must be independent of any factors W which precede X, i.e.,
$$I(Z, W) \quad \forall \quad W : t_W < t_X \qquad (1)$$
2. **Control**: For Z to effect changes in Y (via X) we require that Z and Y be dependent, written:
$$\neg I(Z, Y) \qquad (2)$$
3. **Locality**: To ensure that Z interferes "only" with X, i.e., that its entire effect on Y is mediated by X, we use the conditional independence assertion:
$$I(Z, Y|X) \qquad (3)$$
to read "Z is independent of Y, given X".

Note that (1) and (2) imply (by the axioms of conditional independence [Pearl, 1988b]) that X and Y are dependent, namely, $\neg I(X,\ Y)$.

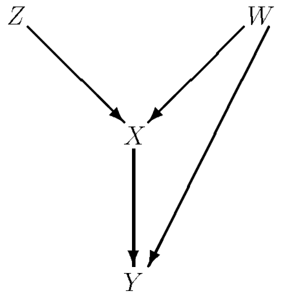

Figure 2

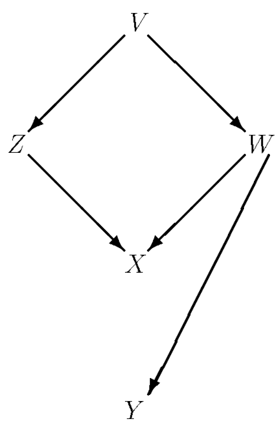

Figure 3

Conditions (1) through (3) constitute the traditional premises behind controlled statistical experiments, with (1) reflecting the requirement that units selected for the expirement be chosen at random from the population under study. They guarantee that any dependency observed between X and Y cannot be explained away by holding fixed some factor W preceding X, hence it must be attributed to genuine causation. The sufficiency of these premises is clearly not a theorem of probability theory, as it relies on temporal relationships among the variables. However, it can be derived from probability theory together with Reichenbach's principle [Reichenbach, 1956], stating that every dependence $\neg I(X, Y)$ requires a causal explanation, namely either one of the variables causes the other, or there must be a variable W preceding X and Y such that $I(X, Y|W)$ (see Figure 2). Indeed, if there is no back path from Z to Y through W (Eq. (1)) and no direct path from Z to Y avoiding X (Eq. (3)) then there must be a causal path from X to Y that is responsible for the dependence in Eq. (2)[9].

In non-experimental situations it is not practical to detach X completely from its natural surrounding and to subject it to the exclusive control of an exogenous (and randomized) variable Z. Instead, one could view some of X's natural causes as "virtual controls" and, provided certain conditions are met, use the latter to reveal non-spurious causal relationship between X and Y. In so doing we compromise, of course, condition (1), because we can no longer guarantee that those natural causes of X are not themselves affected by other causes which, in turn, might influence Y (see Figure 3). However, it turns out that for stable distributions, conditions (2) and (3) are sufficient to guarantee that the association between X and Y is non-spurious, thus justifying Definition 14 for genuine causation.

The intuition goes as follows (see Figure 3): If the de-

[9] Cartwright [Cartwright, 1989] offers a sufficiency proof in the context of linear models.

pendency between Z and Y (and similarly, between X and Y) is spurious, namely, X and Y are merely manifestations of some common cause W, there is no reason then for X to screen-off Y from Z, and condition (2) should be violated. In case condition (2) is accidentally satisfied by some strange combination of parameters, it is bound to be "unstable", as it will be perturbed with any slight change of expiremental conditions.

Conditions (2) and (3) are identical to those in Definition 14, save for the context S which is common to both. The inclusion of the fixed context S is legitimized by noting that if $P(X, Y, Z)$ is a marginal of a stable distribution, then so is the conditional distribution $P(X, Y, Z|S = s)$, as long as S corresponds to variables which precede X.

Definition 14 constitutes an alternative way of recovering causal structures, more flexible than the IC-algorithm; we search the data for three variables Z, X, Y (in this temporal order) that satisfy the two conditions in some context $S = s$, and when such a triple is found, X is proclaimed to have a genuine causal influence on Y. Clearly, permitting an arbitrary context S increases the number of genuine causal influences that can be identified in any given data; marginal independencies and even 1-place conditional independencies are rare phenomenon.

Note that failing to satisfy the test for genuine causation does not mean that such relationship is necessarily absent between the quantities under study. Rather, it means that the data available cannot substantiate the claim of genuine causation. To further test such claims one may need to either conduct experimental studies, or consult a richer data set where virtual control variables are found.

In testing this modeling scheme on real life data, we have examined the observations reported in Sewal Wright's seminal paper "Corn and Hog Correlations" [Wright, 1925]. As expected, corn-price (X) can clearly be identified as a cause of hog-price (Y), not the other way around. The reason lies in the existence of the variable corn-crop (Z) that, by satisfying the conditions of Definition 14 (with $S = \emptyset$), acts as a virtual control of X (see Figure 2). To test for the possibility of reciprocal causation, one can try to find a virtual controller for Y, for example, the amount of hog-breeding (Z'). However, it turns out that Z' is not screened off from X by Y (possibly because corn prices exert direct influence over farmer's decision to breed more hogs), hence, failing condition 3, Y disqualifies as a genuine cause of X. Such distinctions are important to policy makers in deciding, for example, which commodity, corn or hog, should be subsidized or taxed.

8 Non-Temporal Causation and Statistical Time

When temporal information is unavailable the condition that Z precede X (Definition 14) cannot be tested directly and must be replaced by an equivalent condition, based on dependence information. As it turns out, the only reason we had to require that Z precede X is to rule out the possibility that Z is a causal consequence of X; if it were a consequence of X then the dependency between Z and Y could easily be explained away be a common cause W of X and Y (see Figure 2).

The information that permits us to conclude that one variable is not a causal consequence of another comes in the form of an "intransitive triplet", such as the variables a, b and c in Figure 1(a) satisfying: $I(a, b)$, $\neg I(a, c)$ and $\neg I(b, c)$. 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, (recall that in stable distributions, common causes induce dependence among their effects); 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. It is also this intransitive pattern which is used to ensure that X is not a consequence of Y (in Definition 11) and that Z is not a consequence of X (in Definition 12). In definition 14 we have two intransitive triplets, (Z_1, X, Y) and (X, Y, Z_2), thus ruling out direct causal influence between X and Y, implying spurious associations as the only explanation for their dependence.

This interpretation of the intransitive triple is in line with the "virtual control" view of causation. For example, one of the reasons people insist that the rain causes the grass to become wet, and not the other way around, is that they can 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 [Pearl, 1988a, p. 396].

Determining the direction of causal influences from nontemporal data raises some interesting philosophical questions about the nature of time and causal explanations. For example, can the orientation assigned to the arrow $X \rightarrow Y$ in Definition 14 ever clash with temporal information (say by a subsequent discovery that Y precedes X)? Alternatively, since the rationale behind Definition 14 is based on strong intuitions about how causal influences should behave (statistically), it is apparent that such clashes, if they occur, are rather rare. The question arises then, why? Why should orientations determined solely by statistical dependencies have anything to do with the flow of time?

In human discourse, causal explanations indeed carry two connotations, temporal and statistical. The temporal aspect is represented by the convention that a cause should precede its effect. The statistical aspect expects causal explanations (once accounted for) to screen off their effects, i.e., render their effects conditionally independent[10]. More generally, causal explanations are expected to obey many of the rules that govern paths in a directed acyclic graphs (e.g., the intransitive triplet criterion for potential causation, Section 7). This leads to the observation that, if agreement is to hold between the temporal and statistical aspects of causation, natural statistical phenomena must exhibit some basic temporal bias. Indeed, we often encounter phenomenon where knowledge of a present state renders the variables of the future state conditionally independent (e.g., multi-variables economic time series as in Eq. (4) below). We rarely find the converse phenomenon, where knowledge of the present state would render the components of the past state conditionally independent. The question arises whether there is any compelling reason for this temporal bias.

A convenient way to articulate this bias is through the notion of "Statistical Time".

Definition 16 (Statistical Time) *Given an empirical distribution P, a statistical time of P is any ordering of the variables that agrees with at least one minimal causal model consistent with P.*

We see, for example, that a scalar Markov-chain process has many statistical times; one coinciding with the physical time, one opposite to it and the others correspond to any time ordering of the variables away from some chosen variable. On the other hand a process governed by two coupled Markov chains,

$$\begin{aligned} X_t &= \alpha X_{t-1} + \beta Y_{t-1} + \xi_t \\ Y_t &= \gamma X_{t-1} + \delta Y_{t-1} + \xi'_t, \end{aligned} \tag{4}$$

has only one statistical time – the one coinciding with the physical time[11]. Indeed, running the IC-algorithm on samples taken from such a process, while suppressing all temporal information, quickly identifies the components of X_{t-1} and Y_{t-1} as genuine causes of X_t and Y_t. This can be seen from Definition 11, where X_{t-2} qualifies as a potential cause of X_{t-1} using $Z = Y_{t-2}$ and $S = \{X_{t-3}, Y_{t-3}\}$, and Definition 12, where X_{t-1} qualifies as a genuine cause of X_t using $Z = X_{t-2}$ and $S = \{Y_{t-1}\}$ of X_t.

The temporal bias postulated earlier can be expressed as follows:

Conjecture 1 (Temporal Bias) *In most natural phenomenon, the physical time coincides with at least one statistical time.*

Reichenbach [Reichenbach, 1956] attributed the asymmetry associated with his conjunctive fork to the second law of thermodynamics. We are not sure at this point whether the second law can provide a full account of the temporal bias as defined above, since the influence of the external noise ξ_t and ξ'_t renders the process in (4) nonconservative[12]. What is clear, however, is that the temporal bias is *language dependent*. For example, expressing Eq.(4) in a different coordinate system (say, using a unitary transformation $(X'_t, Y'_t) = U(X_t, Y_t)$), it is possible to make the statistical time (in the (X', Y') representation) run contrary to the physical time. This suggests that the apparent agreement between the physical and statistical times is a byproduct of human choice of linguistic primitives and, moreover, that the choice is compelled by a survival pressure to facilitate predictions at the expense of diagnosis and planning.

9 Conclusions

The theory presented in this paper should dispel the belief that statistical analysis can never distinguish genuine causation from spurious covariation. This belief, shaped and nurtured by generations of statisticians [Fisher, 1953, Keynes, 1939, Ling, 1983, Niles, 1922] has been a major hindrance in the way of developing a satisfactory, non-circular account of causation. In the words of Gardenfords [Gardenfors, 1988, page 193]:

> In order to distinguish genuine from spurious causes, we must already know the causally relevant background factors. ... Further, the extra amount of information is substantial: In order to determine whether C is a cause of E, *all* causally relevant background factors must be available. It seems clear that we

[10] This principle, known as Reichenbach's "conjunctive fork" or "common-cause" criterion [Reichenbach, 1956, Suppes and Zaniotti, 1981] has been criticized by Salmon [Salmon, 1984], who showed that some events would qualify as causal explanations though they fail to meet Reichenbach's criterion. Salmon admits, however, that when a conjunctive forks does occur, the screening off variable is expected to be the cause of the other two, not the effect [Salmon, 1984, p. 167]. He notes that it is difficult to find physically meaningful examples where a response variable renders its two causes conditionally independent (although this would not violate any axiom of probability theory). This asymmetry is further evidence that humans tend to reject causal theories that yield unstable distributions.

[11] ξ_t and ξ'_t are assumed to be two independent, white noise time series. Also $\alpha \neq \delta$ and $\gamma \neq \beta$.

[12] We are grateful to Seth Lloyd for this observation.

> often have determinate beliefs about causal relations between events, even if we do not know exactly which factors are causally relevant to the events in question[13].

This paper shows that such extra information is often unnecessary: Under the assumptions of model-minimality (and/or stability), there are patterns of dependencies that should be sufficient to uncover genuine causal relationships. These relationships cannot be attributed to hidden causes lest we violate one of the basic maxims of scientific methodology: the semantical version of Occam's razor. Adherence to this maxim explains why humans reach consensus regarding the directionality and nonspuriousness of causal relationships, in the face of opposing alternatives, perfectly consistent with experience. Echoing Cartwright [Cartwright, 1989] we summarize our claim with the slogan "No Causes In, Some Causes Out".

From a methodological viewpoint, our theory should settle some of the on going disputes regarding the validity of path-analytic approaches to causal modeling in the social sciences [Freedman, 1987, Ling, 1983]. It shows that the basic philosophy governing path-analytic methods is legitimate, faithfully adhering to the traditional norms of scientific investigation. At the same time our results also explicate the assumptions upon which these methods are based, and the conditions that must be fulfilled before claims made by these methods can be accepted. Specifically, our analysis makes it clear that causal modeling must begin with *vanishing* (*conditional*) *dependencies* (i.e. missing links in their graphical representations). Models that embody no vanishing dependencies contain no virtual control variables, hence, the causal component of their claims cannot be substantiated by observational studies. With such models, the data can be used only for estimating the parameters of the causal links once we are absolutely sure of the causal structure, but the structure itself, and especially the directionality of the links, cannot be inferred from the data. Unfortunately, such models are often employed in the social and behavioral sciences e.g. [Kenny, 1979].

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 tens of variables require less than 5000 samples to have their structure recovered by the algorithm. For example, 1000 samples taken from the process shown in Eq. (5), each containing ten successive X,Y pairs, were sufficient for recovering its double-chain structure (and the correct direction of time). The greater the noise, the quicker the recovery.

[13] See also Cartwright [Cartwright, 1989] for a similar position, and for a survey of the literature.

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 and non-spuriousness can only be determined by controlled, manipulative experiments.

It should also be interesting to explore how the new criteria for causation could benefit current research in machine learning. In some sense, our method resembles a search through elements of a version space [Mitchell, 1982], where each hypothesis stands for a causal theory. Unfortunately, this is where the resemblance ends. The prevailing paradigm in the machine learning literature has been to define each hypothesis (or theory, or concept) as a subset of observable instances; once we observe the entire extension of this subset, the hypothesis is defined unambiguously. This is not the case in causal modeling. Even if the training sample exhausts the hypothesis subset (in our case, this corresponds to observing P precisely), we are still left with a vast number of equivalent causal theories, each stipulating a drastically different set of causal claims. Fitness to data, therefore, is an insufficient criterion for validating causal theories. Whereas in traditional learning tasks we attempt to generalize from one set of instances to another, the causal modeling task is to generalize from behavior under one set of conditions to behavior under another set. Causal models should therefore be chosen by a criterion that challenges their stability against changing conditions, and these show up in the data in the form of virtual control variables. Thus, the dependence patterns identified by definition 11 through 14 constitute islands of stability as well as virtual validation tests for causal models. It would be interesting to examine whether these criteria, when incorporated into existing machine learning programs would improve the stability of theories discovered by such programs.

Acknowledgement

We are grateful to Clark Glymour for posing the problem of equivalence in latent structures. Some of the problems treated in this paper were independently explored by Glymour, Spirtes and Schienes [Spirtes et al., 1989, Spirtes and Glymour, 1991], and we thank them for sharing this information with us. Discussions and correspondence with P. Bentler, D. Geiger, C. Granger, M. Hanssens, J. de Leeuw, S. Lloyd, R. Otte, A. Paz, B. Skyrms and P. Suppes are greatly appreciated.

References

[Bobrow, 1985] Bobrow, D. (1985). *Qualitative Rea-*

soning about Physical Systems. MIT Press, Cambridge, MA.

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

[Cliff, 1983] Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate behavioral research*, 18:115 – 126.

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

[Dechter and Pearl, 1990] Dechter, R. and Pearl, J. (1990). Directional constraint networks: A relational framework for causal modeling. Technical Report R-153, UCLA Cognitive Systems Laboratory.

[Eells and Sober, 1983] Eells, E. and Sober, E. (1983). Probabilistic causality. *Philosophy of Science*, 50:35 – 57.

[Fisher, 1953] Fisher, R. A. (1953). *Design of Experiments*. Oliver and Boyd, London.

[Forbus and Gentner, 1986] Forbus, K. D. and Gentner, D. (1986). Causal reasoning about quantities. *Proceedings Cognitive Science Society*, pages 196 – 207.

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

[Frydenberg, 1989] Frydenberg, M. (1989). The chain graph markov property. Technical Report 186, Department of Theoretical Statistics, University of Aarhus, Denmark.

[Gardenfors, 1988] Gardenfors, P. (1988). Causation and the dynamics of belief. In Harper, W. and Skyrms, B., editors, *Causation in Decision, Belief Change and Statistics II*, pages 85 – 104. Kluwer Academic Publishers.

[Geffner, 1989] Geffner, H. (1989). *Default Reasoning: Causal and Conditional Theories*. PhD thesis, UCLA Computer Science Department, Los Angeles, CA.

[Geiger et al., 1990] Geiger, D., Paz, A., and Pearl, J. (1990). Learning causal trees from dependence information. In *Proceedings, AAAI-90*, pages 770 – 776, Boston, MA.

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

[Good, 1983] Good, I. J. (1983). A causal calculus. *British Journal for Philosophy of Science*, 11 and 12 and 13:305 – 328 and 43 – 51 and 88. reprinted as Ch. 21 in Good Thinking University of Minnesota Press, Minneapolis, MN.

[Granger, 1988] Granger, C. W. J. (1988). Causality testing in a decision science. In Harper, W. and Skyrms, B., editors, *Causation in Decision, Belief Change and Statistics I*, pages 1 – 20. Kluwer Academic Publishers.

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

[Iwasaki and Simon, 1986] Iwasaki, Y. and Simon, H. A. (1986). Causality in device behavior. *Artificial Intelligence*, 29(1):3 – 32.

[Kautz, 1987] Kautz, H. (1987). *A formal Theory of Plan Recognition*. PhD thesis, University of Rochester, Rochester, N.Y.

[Kenny, 1979] Kenny, D. A. (1979). *Correlation and Causality*. Wiley, New York.

[Keynes, 1939] Keynes, J. M. (1939). Professor tinbergen's method. *Economic Journal*, 49:560.

[Lifschitz, 1987] Lifschitz, V. (1987). Formal theories of action. In *Workshop of the Frame Problem in AI*, pages 35 – 57, Kansas.

[Ling, 1983] Ling, R. (1983). Review of "Correlation and Causation" by D. Kenny. *Journal of the American Statistical Association*, pages 489 – 491.

[Mitchell, 1982] Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, 18:203 – 226.

[Niles, 1922] Niles, H. E. (1922). Correlation, causation, and Wright theory of "path coefficients". *Genetics*, 7:258 – 273.

[Patil et al., 1982] Patil, R. S., Szolovitz, P., and Schwartz, W. B. (1982). Causal understanding of patient illness in patient diagnosis. In *Proceedings of AAAI-82*, pages 345 – 348.

[Pearl, 1978] Pearl, J. (1978). On the connection between the complexity and credibility of inferred models. *International Journal of General Systems*, 4:255 – 264.

[Pearl, 1988a] Pearl, J. (1988a). Embracing causality in formal reasoning. *Artificial Intelligence*, 35(2):259 – 71.

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

[Pearl, 1990] Pearl, J. (1990). Probabilistic and qualitative abduction. In *Proceedings of AAAI Spring Symposium on Abduction*, pages 155 – 158, Stanford.

[Pearl et al., 1989] Pearl, J., Geiger, D., and Verma, T. S. (1989). The logic of influence diagrams. In Oliver, R. M. and Smith, J. Q., editors, *Influence Diagrams, Belief Networks and Decision Analysis*, pages 67 – 87. John Wiley and Sons, Ltd., Sussex, England.

[Popper, 1959] Popper, K. R. (1959). *The Logic of Scientific Discovery*. Basic Books, New York.

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

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

[Rubin, 1989] Rubin, H. (1989). Discussion of "The Logic of Influence Diagrams" by Pearl at al. In Oliver, R. M. and Smith, J. Q., editors, *Influence Diagrams, Belief Networks and Decision Analysis*, pages 83 – 85. John Wiley and Sons, Ltd., Sussex, England.

[Salmon, 1984] Salmon, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press., Princeton.

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

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

[Skyrms, 1986] Skyrms, B. (1986). *Causal Necessity.* Yale University Press, New Haven, CT.

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

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

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

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

[Suppes and Zaniotti, 1981] Suppes, P. and Zaniotti, M. (1981). When are probabilistic explanations possible? *Synthese*, 48:191 – 199.

[Verma, 1991] Verma, T. S. (1991). Invariant properties of causal models. Technical report, UCLA Cognitive Systems Laboratory.

[Verma and Pearl, 1990] Verma, T. S. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings 6th Conference on Uncertainty in AI*, pages 220 – 227.

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

[Wright, 1925] Wright, S. (1925). Corn and hog correlations. Technical Report 1300, U.S. Department of Agriculture.