Whether Non-Correlation Implies Non-Causation

Eric Neufeld and Sonje Kristtorn
Department of Computer Science, 110 Science Place
University of Saskatchewan, Saskatoon, Canada S7N 5C9
eric@cs.usask.ca, wlf323@cs.usask.ca

Abstract

It has been well argued that correlation does not imply causation. Is the converse true: does non-correlation imply non-causation, or more plainly, does causation imply correlation? Here we argue that this is a useful intuition of the semantic essence of the faithfulness assumption of causal graphs. Although the statement is intuitively reasonable, it is not categorically true (but it is true with probability one), and this brings into question the validity of causal graphs. This work reviews Cartwright's arguments against faithfulness and presents a philosophical case in favor of the faithfulness assumption. This work also shows how the causal graph formalism can be used to troubleshoot scenarios where faithfulness is violated.

Introduction

Causal graphs have been studied as such for more than a decade. Originally introduced as Bayesian nets (Pearl 1988), they demonstrated the practicality of purely probabilistic reasoning to an AI community that believed probability theory was epistemologically inadequate even for mundane knowledge. Causal graphs gave both a compact representation of joint distributions of many variables, and sound and efficient inference algorithms.

At that time, many groups sought alternatives to probability for the purpose of representing uncertain knowledge. Several groups pursued nonnumeric, or symbolic alternatives, such as endorsements or the various default logics. (See (Kanal and Lemmer 1986) for discussions.) The reasons for different strategies varied, but it would be fair to say that many believed that intelligent agents were capable of sophisticated reasoning strategies without numeric information, and furthermore, that accurate statistics were rarely available. This is obviously true in commonsense domains, but even in areas such as medical diagnosis, some diseases evolve so quickly that it is difficult to collect accurate statistics before the population changes significantly.

Other groups, citing limitations of the expressive power of traditional probability, pursued alternate numeric calculi. This group formed the core of the early UAI (Uncertainty in Artificial Intelligence) community (Kanal and Lemmer 1986). Three that became prominent were Certainty Factors (the calculus driving expert systems like MYCIN), Fuzzy Logic, and belief functions (also known as Dempster-Shafer theory). Early papers on causal graphs also appeared at this time. The result was a thorough discussion of the foundations of uncertain reasoning.

Initially, the 'causal' aspect of causal graphs was informal. In simple settings such as diagnostic bipartite causal graphs (e.g., the work of Peng and Reggia (1986) translated to causal graphs), this was abundantly clear. Root nodes were unconditionally independent diseases, leaf nodes were conditionally independent symptoms, and arrows pointed in the direction of causality. Causal graphs were found to have many useful qualitative interpretations. Pearl and Verma (1991) offered probabilistic definitions of potential and genuine causality that not only gave philosophical justification for this phenomenon, but also offered an algorithm for recovering causal structure from sufficiently rich numeric data. The algorithm could also identify certain spurious associations, that is, variables whose correlation was due to both variables being caused by a hidden unmeasured variable.

The semantic correctness of Pearl and Verma's inductive causation (IC) algorithm, rested on two important assumptions. The first was the availability of exact distributions. The second was the faithfulness (SGS 1993) assumption (also known as stability (Pearl 2000)), which, "conveys the assumption that all the independencies embedded in [the probability distribution] P are stable, that is, they are entailed by the structure of the model D and remain invariant to any change in the parameters Θ_D" (Pearl 2000). The parameters are the numerical conditional probabilities stored at each node.

Pearl justifies the faithfulness (stability) assumption with a visual analogy. Suppose we see a picture of a chair, and need to decide between theory T_1, which states the object in the picture is a chair, and T_2, which states the picture contains two chairs aligned so one hides the other. T_1 is invariant to the angle of view, and T_2 is unlikely. In this sense T_1 is simpler than T_2.

Theoretical counterexamples to the faithfulness assumption exist. The construction is easy and similar to Simpson's paradox, and moreover, the counterexamples can plausibly occur in real life. However, even though the counterexamples are theoretically unlikely (using a Bayesian analysis), the possibility is as troubling as Simpson's paradox.

In the sequel, we review terminology for causal graphs, causal terminology and algorithms for the construction of causal graphs from data. We review concerns about the faithfulness assumption, but argue that the intuitions are valid and useful. Moreover, in the event of a counterexample, we show how the formalism itself can be used to troubleshoot the model.

Notation and terminology

A causal graph is a pair *(D,P)*, where *D=(V,E)* is a directed graph, where *V* is a set of variables and *E* is a set of directed edges (arcs) between variables. *P* is a probability distribution over the variables in *V*. For any variable *A*, *parents(A)* denotes the parents of *A*, the direct predecessors of *A*, or the direct causes of *A*. Associated with *A* is a local distribution *f(A, parents(A))*, which gives a distribution for *A* for any set of values that the parents(*A*) take on. Moreover, the distribution *P* can be decomposed into these independent conditional distributions at the nodes. Commonly *f* is the familiar discrete conditional probability distribution. For the present, it suffices to say that the formalism generalizes beyond discrete and Gaussian variables (in the form of structure equations, or path analysis), but for this discussion, we use discrete distributions in the text and a Gaussian example in images.

For discrete variables, the graph's structure encodes the information that *P* factors into the product of the conditional distributions stored at the nodes. That is,

$$P(v_0,\ldots,v_n) = \Pi_i P(v_i|parents(v_i)). \qquad (1)$$

Variables *A, B* are conditionally independent given a set of variables *C* if

$$P(A,B|C) = P(A|C)\cdot P(B|C)$$

for all outcomes of *A, B* and *C*. If *C* is empty, then *A,B* are unconditionally independent. The factorization in Equation 1 has the property that vertices obey the *Markov Condition*, that a vertex is conditionally independent of its non-descendants given its parents. In the setting of a causal graph, the Markov condition implies many other conditional independencies. These can be detected from *D* alone using a graph-theoretic criterion called *d-separation*. (Pearl 2000).

Definition 1. (Potential Cause) (Pearl and Verma, 1991) *A* is a potential cause of *C* if there is a variable *B* and a context (set of variables) *S* such that

i. *A, B* are independent given *S*,
ii. there is no set *D* such that *A,C* are conditionally independent given *D,* and
iii. *B, C* are dependent.

A simplified version of Pearl's causal graph construction (IC-algorithm) follows.

1. For each pair of vertices *A,B* in *V*, search for a subset *S* of *V* (including the empty set) such that *A* is conditionally independent of *B* given *S*. If no such *S* exists, add an undirected edge between *A* and *B*.
2. For each collinear triple of vertices A—C—B, where *A* and *B* are not directly connected, test whether *A, B* and *C* satisfy the relationship of potential cause as in Definition 1. If yes, add head-to-head arrows at *C*. Repeat this step until no more arrows can be added. (It is acceptable to obtain double-ended arrows, but this is not discussed here.)
3. For all remaining undirected arcs, add arrows, but do not create any new structures like those in Step 2, or directed cycles. Given these constraints, certain arcs may be oriented in either direction. Such arcs should be left undirected.

The output of this algorithm is a graph containing undirected arcs, directed arcs (arrows), and arcs directed in both directions. Directed arcs indicate causal links. Arcs directed at both ends indicate the location of hidden causal variables. Undirected arcs indicate insufficient information to orient an arc.

When constructing the directed causal graph from data, this algorithm twice makes critical use of the faithfulness assumption. In Step 1, it places an arc between two nodes, only if no set *S* makes them independent. Under the usual interpretation of faithfulness, the absence of any form of independence forces a link to be added, since the only dependencies are those implied by structure.

We argue that the intuition behind faithfulness is the idea that causation implies correlation. Under this interpretation, we think of the IC-algorithm in terms of the complementary action of not adding a link when independence is discovered. All links added to the graph have the potential of being oriented to imply causation. Thus, independence (or the lack of dependence, or non-correlation) in any setting means no causal link is added, or, non-correlation implies non-causation.

The faithfulness assumption is also used in Step 2, embedded in the definition of potential cause. Under the usual interpretation of faithfulness, consider all possible orientations of the arcs connecting *A, B, C* as shown in Figure 1.

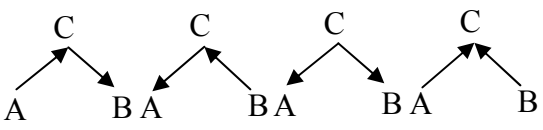

Figure 1

Using the faithfulness assumption in the traditional sense, independencies in the data are implied by structure only, and the head-to-head structure is the only one of the four that implies the data.

Under our interpretation, the assumption that causation implies correlation rules out the first three possibilities, by reasoning as follows. Consider a setting where *A* and *B* are independent, and there is a path from *A* to *B*. If causation implies correlation, the first two orientations are easily ruled out, since *A* and *B* are not correlated. In the third case, one also expects correlations between effects of a common cause. However, it is straightforward to use the factorization properties of a causal graph to verify that only the last graph is always consistent with the definition of potential cause, regardless of assignment of conditional probability distributions to the vertices. That is, the definition of potential cause is based on the assumption that causation implies correlation.

A causal inference algorithm such as this one can't do much without an assumption like this, although it is easy to construct counterexamples that show the faithfulness assumption to be imperfect. Consider a graph such as that in Figure 2(a), below.

(a) **(b)**

Figure 2

Because small (3 node) graphs with no unconditional independence can have three different labelings (that is, the first three labelings in Figure 1), the graph in Figure 2(a) is the smallest graph that the IC-algorithm, or TETRAD (SGS 1993), can orient. However, it is straightforward to construct a graph with the topology of Figure 2(b): just use the graph in Figure 2(a) to compute the conditional probability distributions for each node in the second graph. Although this is a simple exercise, it is interesting for two reasons: 1) it introduces an arc where none existed before, and 2) the causal directions of the original arcs in Figure 2(a) are reversed.

To give this some meaning, we borrow from a real-life example (Dennis et al, 1993). A recent study showed that sunscreen users may be at increased risk of melanoma. For the present discussion, we adjusted the distribution so that sunscreen (*B*) has no effect on melanoma (*A*). However, further study reveals light-skinned people use sunscreen more often than darker-skinner people. Because natural pigmentation (*C*) also offers protection from melanoma, by chance, the inverse relationship between these two causal influences makes the two influences exactly cancel each other out. However, applying the IC-algorithm blindly leads to an incorrect result – that melanoma and sunscreen usage have causal influence on pigmentation, as in Figure 2(b). This construction is similar to, and provides a causal variation of, Simpson's paradox by reversing causal direction. Although this is implausible in this setting, we provide a possible interpretation. The determination of the actual relationships may have consequences for policy makers in public health and for consumers.

How does this impact the enterprise of causal inference from observational data?

Cartwright's critique of faithfulness

A counterexample to a premise is generally certain death for a theory. A single counterexample (and there are infinitely many in the present case) shows that we cannot say that causation implies correlation (in the usual sense of the first order logic), which we have argued above is the intuition of the faithfulness assumption.

However, it is possible to argue that probability of correlation given causation has measure one. For two variables *A* and *C*, for any distribution of *A*, there is exactly one joint distribution of *A, C* such that the two variables are independent, but in all the remaining uncountably many distributions, the two variables are dependent.

This ingenious Bayesian argument is used for the faithfulness assumption, and assumes that all probability distributions are equally likely. This latter assumption runs into difficulty, clearly expressed by Cartwright (1999). She states

> It is not uncommon for advocates of DAG-techniques to argue that cases of cancellation will be extremely rare, rare enough to count as non-existent. That seems to me unlikely, both in the engineered devices that are sometimes used to illustrate the techniques and in the socio-economic and medical cases to which we hope to apply the techniques. For these are cases where means are adjusted to ends and where unwanted side effected tend to be eliminated wherever possibly, either by following an explicit plan or by less systematic fiddling.

Elsewhere (Cartwright, 2003), she goes into considerably more detail. In essence, she states that Pearl's argument puts "structure first", and parameters second. However, she claims that one cannot have one without the other, and gives a convincing example. Birth-control pills may cause thrombosis, and thus we try to weaken the strength with which they so do. Thus, she concludes "Getting the cancellation that stability/faithfulness prohibits is important to us". More generally, she argues, that probability and causal structures constrain each other. If the probabilities are fixed, then we are constrained from building certain causal structures, or, (in the case of faithfulness), vice-versa.

It may be easier in applied science and/or engineering to design a counterprocess to cancel certain effects of a process than it is to eliminate the process causing the effect in the first place. Thus, a case can be made that Nature, as engineer, frequently uses the same ploy. For example, we talk about the balance of nature. Since the theory of causal graphs cannot get far without the faithfulness assumption, this is potentially devastating.

Revisiting the sunscreen example, this provides an unusual meta-interpretation for Figure 2(a). Perhaps sunscreen usage and melanoma independently conspire to eliminate the causal power of skin pigmentation. This doesn't seem plausible. However, in a world where we don't know the true number of causal influences on any effect, it seems possible, if experimental error is considered, that some pair of them might almost exactly cancel, and that we might have unfortunately picked that pair.

Whether causation implies correlation

An assumption that precedes the faithfulness assumption is an assumption regarding the existence of causality. Karl Pearson saw cause and effect as an old and simplistic concept, as extreme cases on a continuum of dependence, rendered obsolete by the new ideas of association. Cartwright's view suggests that the new theories of causation seem to be too accident-prone to be trusted.

This is our view: The world we inhabit cannot be experienced directly. It is knowable only through our senses, which vary widely among individuals, and measurable only through devices constructed in that same world that are subject to some error. However, simple direct cause and effect relationships may properly exist in this world, and may be measurable in some sense, but the outdegree and indegree of every node may be prohibitively high. If causation doesn't properly exist in the real world, we could, as in other mathematical sciences, use causality as an idealization from which to make predictions about interventions.

In this ideal world, causality exists. In this ideal world, causality implies correlation, almost always. "Almost always" is used in the same sense as elsewhere in mathematics: there are at most countably many exceptions to an uncountable number of truths.

This is qualitatively and philosophically different from the converse idea that correlation implies causation, which is a logical error. Even if we assume that the presence of correlation between two variables implies the presence of causation, the correlation says nothing about the direction of causation, and furthermore, correlation can be explained by both variables being caused by some third variable. The next problem is measurement (Kyburg 1984).

In the case of discrete variables, we run into a host of epistemological problems. The very first is that of natural kinds. If we want to measure the proportion of birds that fly, we need a criterion to determine to distinguish birds from non-birds. As well, we need a definition of what it means to be able to fly. Even if this is possible, there is the problem of knowing when we have found all the entities we wish to call birds. There is also the fact that world doesn't stay still while we are counting: as Bacchus (1989) observes, we might do our counting in the spring, when most birds are flightless nestlings.

These problems get even more complicated once we try to measure continuous variables, whether with sticks and springs or laser beams. A practical theory of causation must address these questions of measurement, even if the ideas are theoretically robust.

However, to get any data in the first place (apart from ideally generated distributions), we must accept statistical, or *measurement*, correlation as a proxy for *actual* correlation. This may seem like a sleight-of-hand, but other methods for reliable causal inferences are subject to practical problems. For instance, Robins and Wasserman (1999), state that

> We are not claiming that inferring causal relationships empirically is impossible. Randomized studies with complete compliance are a well-known example where reliable causal inference is possible.

The many practical difficulties inherent in collecting samples for studies, and many practical difficulties regarding compliance, in the view of Robins and Wasserman, do not undermine the idea of randomized studies enough to abandon the idea. (In fact, much simpler statistical inferences are subject to practical problems.)

A counterargument is that one may inspect samples and discard them if they are found not to be representative, and one may observe noncompliance. However, suppose the measured correlations are close to correct, but exist for the kind of reasons (i.e., cancellation) Cartwright posits. If that is the case, we still have a means for testing the predictive power of the theory. This is illustrated in the next section with an example exploring the sunscreen example using a causal visualization tool developed in our group.

Testing causal theories

A useful feature of causal theories is that they give us, under certain assumptions, the computational power to distinguish between *seeing* and *setting*. (Freedman (1997) uses the terms *observation* and *intervention*.) *Seeing* is the common procedure of computing a conditional expectation—given a subpopulation of a known population, what is the posterior distribution of all related variables in the subpopulation? For example, we may wish

to compute $p(M|see(S))$, the probability that someone we *see* using sunscreen might develop melanoma. This can be computed as the ordinary conditional probability $P(M|S)$.

Setting is about the consequences of actions (or interventions) and requires a subtler calculation. An individual wants to know the net effect of using sunscreen, for that individual, whether sunscreen use decreases or increases the overall probability of melanoma, in light of conflicting opinions. These two are computed as follows (Pearl, 2000):

$$P(M|see(S)) = p(M|DS)p(D|S) + p(M|\neg DS)p(\neg D|S),$$
$$P(M|set(S)) = p(M|DS)\,p(D) + p(M|\neg DS)p(\neg D).$$

Roughly, to compute the effect of setting S, we assume that the effect of sunscreen among all the Ds will have the same effect as it does among those who currently use sunscreen. (This explains the first multiplicative term in the second equation.) We make the same assumption for the $\neg D$s. Although this formulation goes back to Stotz and Wold (1960), it has a simple implementation in causal graphs: erase arcs incoming to the *set* variable, and change all distributions involving the *set* variable accordingly. For correctness, see Pearl (2000).

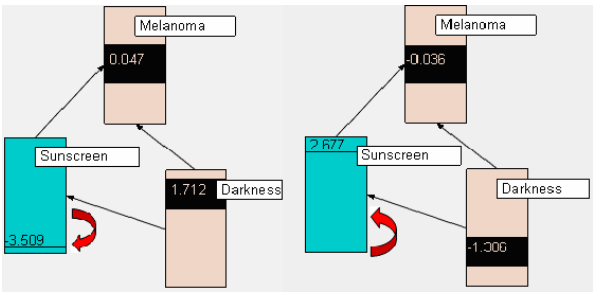

Figure 3. Melanoma does not respond much to changes in sunscreen. Added red arrows indicate mouse movement during interaction

Suppose the resulting model is incorrect as a consequence of the case that causal relationships cancel each other out, showing *melanoma* and *sunscreen* as independent common causes of *darkness* (pigmentation) of skin, and mathematically is a minimal representation of possible causal relationships in a world coherent with the raw data. The user intuits this is incorrect and eventually constructs the correct model of causal relationships.

Figure 3 shows the results of a user exploring *see*ing using a visualization tool we have developed (Neufeld *et al*, 2005). The user grabs the value of the *sunscreen* variable, drags it up and down, and finds that *sunscreen* and *melanoma* are unrelated, as found by the original data mining tool (e.g., TETRAD). The statistical explanation is that *darkness* acts as a suppressor variable or confound. Light-skinned individuals are more likely to use sunscreen than dark-skinned people. However, they are also more likely to develop melanoma, exactly canceling the effect of the sunscreen.

An experienced data analyst would revisit the data and ask what happens when *sunscreen* is manipulated after *darkness* is first fixed at some value Figure 4 illustrates what needs to be done. The user first chooses to *see* a fixed value for the *darkness* variable, and then *sees* a range of values for *sunscreen by* dragging it up and down. The pairs of before and after images in Figure 4 now reveal the correct relationship between *sunscreen* and *melanoma*. Whether *darkness* is high or low, fixing *darkness,* and then increasing *sunscreen* results in a decrease to *melanoma.*

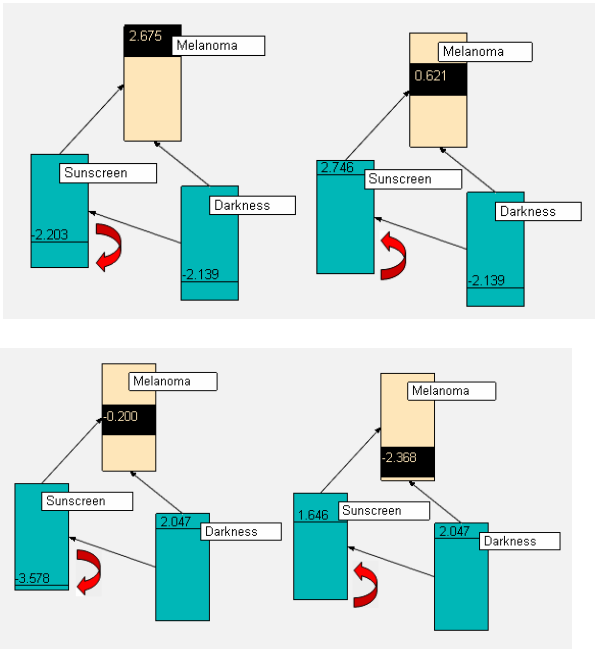

Figure 4. Melanoma responds to seeing changes in sunscreen within a skin type. The top pair of graphs shows the change in melanoma with sunscreen for light-skinned persons, and the bottom pair shows melanoma change for dark-skinned persons. Ranges differ, but melanoma incidence consistently decreases with increased sunscreen usage.

The user needs to perform a double *seeing* operation because there are two paths of probabilistic influence from *sunscreen* to *melanoma* that cancel each other out.

Because there are few potential confounds in this three node world, trying all *see* operations is not logistically difficult. However, in a richer dataset, this process may be cumbersome. *Setting* summarizes this combination of actions, as shown in Figure 5.

Moreover, the predictions from the hypothesized cause-effect relationships give us something we can plausibly check, if not in the real world, in the idealized world.

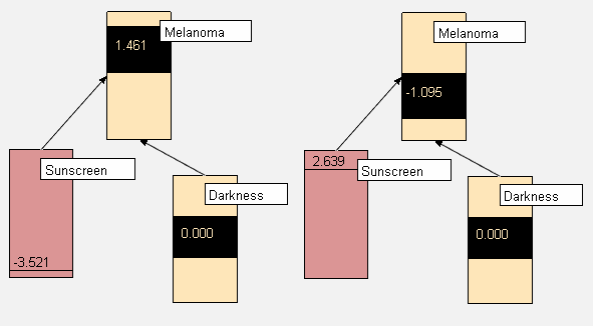

Figure 5. Melanoma responds appropriately to setting of the sunscreen variable.

Conclusions and ongoing work

Faithfulness is presently defined as the assumption that all independencies are "true" independencies, that is, probabilistic independencies are causal independencies. We have replied to a criticism of faithfulness by suggesting that the intuitive semantic content of this assumption is that causation implies correlation. This idealization, which is almost always true, provides a basis for constructing and testing causal theories. We are exploring other characterizations. We have shown how predictions of such a theory can be explored, using a visualization tool that acts as a cognitive prosthetic (Neufeld *et al*, 2003). These predictions can be tested either against our intuitions or with experiments or further measurements, and the results can be used to further constrain the data mining tool. In ongoing work, we are looking both at improving the visualization tool, and at other characterizations and criticisms of faithfulness.

Acknowledgments

This research was support by a grant from the Natural Science and Engineering Research Council of Canada. The comments of some observant reviewers improved the presentation.

References

Bacchus, F. (1989) A Modest, but Semantically Well Founded, Inheritance Reasoner. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* Detroit, 1104-1109

Cartwright, N. (1999) The Dappled World: A Study of the Boundaries of Science. Cambridge University Press

Cartwright, N. (2003) What is Wrong with Bayes Nets? In *Probability is the Very Guide of Life*. Henry E. Kyburg and Mariam Thalos, Eds. OpenCourt Chicago

Dennis L.K, Beane Freeman L.E., VanBeek M.J. (2003) Sunscreen use and the risk for melanoma: a quantitative review. Annals of Internal Medicine:139(12):966-78

Freedman, D.A. (1997). From association to causation via regression. In V.R. McKim and S.P. Turner (Eds). Causality in crisis: statistical methods in the search for causal knowledge in the social sciences. South Bend: University of Notre Dame Press, 113-161

Kanal, L. and Lemmer, J., eds. (1986) *Uncertainty in Artificial Intelligence* North Holland Amsterdam

Kyburg, H. E., Jr. (1984) *Theory and Measurement* Cambridge University Press

Neufeld, E., Kristtorn, S., Guan, Q., Sanscartier, M. and Ware, C. (2005) Exploring Causal Influences. In *Proceedings of Visualization and Data Analysis 2005* San Jose, to appear

Neufeld, E., Callele, D., Mould, D., Kristtorn, S., and Rabu, R. A contribution to the theory and practice of cognitive prostheses. In *Proceedings of Smart Graphics 2003*, 241-250

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers Inc. San Francisco, CA

Pearl, J and Verma, T. (1991) A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, San Mateo, pp. 441-452

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

Peng, Y and Reggia, J.A. (1986) Plausibility of diagnostic hypotheses. In *Proceedings of the 5th International Conference on AI,* Philadelphia, 140-145

Robins, J. and Wasserman, L. (1999) On the Impossibility of Inferring Causation without Background Knowledge. In *Computation, Causation and Discovery*, Glymour and Cooper, eds. AAAI Press, California

Spirtes P., Glymour C., & Scheines R. (1993). Causation, Prediction and Search. Lecture Notes in Statistics, vol. 81. New York: Springer-Verlag.

Stotz, R.H. and Wold, H.O.A (1960) Recursive versus nonrecursive systems: An attempt at synthesis. Econometrica 28, 417-427