1-19-1999

The Limits of Causal Knowledge

James M. Robins
Harvard University

Richard Scheines
Carnegie Mellon University

Peter Spirtes
Carnegie Mellon University

Larry Wasserman
Carnegie Mellon University

The Limits of Causal Knowledge

James M. Robins, Richards Scheines
Peter Spirtes and Larry Wasserman

Harvard University and Carnegie Mellon University

January 19, 1999

Technical Report No. CMU-PHIL-97

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

The Limits of Causal Knowledge

James M. Robins, Richard Scheines, Peter Spirtes and Larry Wasserman
Harvard University and Carnegie Mellon University

June 2, 1998

MANUSCRIPT IN PROGRESS

We investigate the asymptotic consistency of causal inference procedures in the framework of directed acyclic graphs (DAG's) as developed by Spirtes, Glymour and Scheines (SGS) and Pearl and Verma (PV). We show that there exist "pointwise consistent" but not "uniformly consistent" procedures. These results have implications for making inferences based on finite sample sizes and for constructing valid confidence intervals for causal effects.

1. INTRODUCTION

The problem of inferring causal relationships between variables is a topic of continuing interest. One promising approach that has received much attention lately is based on directed acyclic graphs (DAG's). In this approach, direct causal relationships among variables are represented by arrows on a DAG. Hence, we refer to this as the "DAG framework" for causal inference. Under weak assumptions, the DAG implies certain conditional independence assumptions among the variables. In practice, one tries to deduce features of the underlying DAG from observed independence relations in the data. Statisticians are usually concerned about the effect of unobserved confounding variables. But the DAG framework explicitly allows for unobserved confounding. How is it possible to infer causal relationships when there are unobserved confounders? This would seem to contradict standard statistical wisdom about the difficulties of inferring causality from observational data. One might suspect that the inferences are possible because the DAG approach has strong hidden assumptions. We show in this paper that the apparent contradiction between the results in the DAG framework and the conventional statistical wisdom are not really at odds with each other. Learning about causal relationships is possible in the DAG framework but the type of learning is based on pointwise consistency, rather than uniform consistency. The goal of this paper is to make this point precise and, in doing so, to cast light on what is and is not learnable in causal inference.

A brief overview of the framework of the paper and the results are as follows. We have a set of random variables $V = (\mathcal{O},\mathcal{U})$ but we only observe the variables $\mathcal{O}$. The other variables $\mathcal{U}$ are latent, unobserved variables. For example, suppose $\mathcal{O} = \{X, Y\}$ where X

is a binary variable that indicates whether or not a subject is a smoker or not, and Y is a binary variable that indicates presence or absence of some lung disease. $\mathcal{U}$ represents all possible unobserved confounding variables that are related to X and Y. We are interested in the causal relationships between these variables in $\mathcal{O}$. For example, we want to know if X causes Y. The possibility hat we are omitting relevant variables that affect both X and Y is obviated since $\mathcal{U}$ is assumed to contain all possible confounding variables.

In the DAG framework, we assume that the variables V have some joint distribution P which is "Markov" and "faithful" with respect to some DAG G. Each vertex in G corresponds to one variable in V and each arrow in G represents a direct causal relationship. The Markov assumption means roughly that the absence of arrows in G induce independence relationships between variables. The faithfulness assumption means roughly that the presence of arrows in G induce dependence relationships between variables. Our goal is to infer features of G such as "is there an arrow from X to Y?" Specifically, consider two questions:

(1) Give only the joint distribution $P_{\mathcal{O}}$ for $\mathcal{O}$, can we deduce certain features of G?

(2) Given n i.i.d. data points from $P_{\mathcal{O}}$, can we reliably infer certain features of G?

Spirtes, Glymour and Scheines (SGS) (1993) and Pearl and Verma (PV) (1991) show that the answer to (1) is yes. More precisely they show that, given $P_{\mathcal{O}}$, it is possible to find a non-trivial set of DAG's that contains the true DAG G. Two paradigmatic examples in Section 2 will show how the reasoning works.

In practice, we only have a sample from $\mathcal{P}_{\mathcal{O}}$. This brings us to question (2). Now the answer to (2) depends on what we mean by "reliable." SGS claim, correctly, that from the sample, we can consistently deduce certain features of the DAG for V. Specifically, let G_0 denote the true DAG and let $\mathcal{O}^n = (\mathcal{O}_1, \ldots, \mathcal{O}_n)$ denote the observed data. SGS create a procedure which, given $\mathcal{O}^n$, yields a a non-trivial set of graphs $\mathcal{G}(\mathcal{O}^n)$. They claim, correctly, that

$$Pr(G_0 \notin \mathcal{G}(\mathcal{O}^n)) \to 0 \quad \text{as } n \to \infty. \tag{1}$$

This may seem surprising since G_0 involves an unknown number of unobservable variables $\mathcal{U}$.

Equation (1) is an example of pointwise consistency: The probability of an incorrect result tends to 0. On the other hand, there are good reasons that one might demand a stronger mode of consistency, namely, some form of uniform consistency. We will show that – in the absence of randomization or specific assumptions about confounding – causal

procedures cannot in general be uniformly consistent.

In Section 2 we give a brief introduction to causal inference based on DAG's. In Section 3 we discuss consistent tests. In Section 4 we take up the main claim of the paper: the non-existence of uniformly consistent causal inference procedures. Section 5 examines the implications for confidence intervals and point estimates. Section 6 discusses sensitivity analysis, section 7 discusses the philosophical implications. Proof of results are in Section 8. Throughout the paper, we write $X \amalg Y$ to mean that the random variables X and Y are independent. Similarly, we write $X \amalg Y|Z$ to mean that the random variables X and Y are independent given Z.

2. CAUSAL INFERENCE

Let $V = (X_1, \ldots, X_k)$ be a set of random variables. For simplicity, assume each X_j is discrete. Consider a DAG G where each node in G corresponds to one variable in V. An arrow from X_i to X_j represents the fact that X_i is a direct cause of X_j. Let PA_j be the parents of X_j, i.e. the set of variables with arrows pointing into X_j. A distribution P with mass function p "is Markov to G" if its density p can be written

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{n} p(x_i|pa_i).$$

Let $\mathcal{P} \equiv \mathcal{P}(G)$ be all probability distributions that are Markov to G. Given $p \in \mathcal{P}$, let $\mathcal{I}(p)$ represent all independence and conditional independence relation implied by p. We say that p is "faithful" to G if

$$\mathcal{I}(p) = \cap_{q \in \mathcal{P}} \mathcal{I}(q).$$

In other words, p is faithful if it does not possess "extra" independence relations not shared by all the others distributions in $\mathcal{P}$. In cases where $\mathcal{P}$ can be parameterized by a family of distributions with a parameter θ of finite dimension, the set of unfaithful distributions typically has Lebesgue measure 0 (reference xxxx). This is one reason for taking faithfulness as an assumption. Moreover, since unfaithfulness represents independencies that arise by coincidence rather than through structural reasons, it is often reasonable to rule out such distributions on the principle of parsimony.

The faithfulness assumption provides an extremely powerful lever for turning judgments about statistical independence into claims about causality. If statistical judgments about

independence are correct, then in some circumstances (which we will illustrate here) faithfulness can eliminate the possibility of unmeasured confounders and establish the existence of a causal relation. Let Ω be all the probability distributions that are Markov and faithful to a DAG G. Then the independence relations in every member of Ω can be computed directly from G by applying Pearl's d-separation criterion to G (Pearl 1988, SGS 1993). Thus it is easy to begin with a DAG over a set of random variables V and compute the independence relations that hold among V in every probability distribution Markov and faithful to G, a set which we say is faithfully generated by G.

Inferring causal relations goes the other way, however. We begin with the independence relations implied by a distribution P, which in practice we must decide statistically from data, and make inferences about the causal structure that might have faithfully generated P. Not surprisingly, many different DAGs over V can faithfully generate the same set of independence relations. Such DAGs are d-separation equivalent, and they are indistinguishable from independence relations alone. For example, illustrating the adage that correlation is not causation, if $V = \{X, Y\}$ and X and Y are not independent, the DAGS $G1 = X \rightarrow Y$, and $G2 = X \leftarrow Y$ are d-separation equivalent and thus indistinguishable from independence facts alone, even assuming faithfulness. If we further allow that our data might only include a proper subset of the variables in the DAG that faithfully generated the full data, i.e., there might be unmeasured confounders, then the situation is even worse. Suppose that $V = \{X, Y, U\}$, but that the measured variables $O = \{X, Y\}$ and $\sim X \amalg Y$ in P_O, the joint distribution over O, then the set of DAGs that could have faithfully generated a distribution P for which P_O is the marginal over O is in Figure RS1 1.

How are we to know that U is the only unmeasured confounder? The DAGs in Figure RS2, and an infinity of others like them, could also have faithfully generated a distribution P for which P_O is the marginal over O.

The situation may appear completely hopeless for causal inference, but isn't. Suppose that X and Y are independent. Then there is no DAG with our without unmeasured variables that could have faithfully generated a distribution P for which P_O is the marginal over O in which:

X is a cause of Y, or

Y is a cause of X, or

there exists some unmeasured U that is a cause of both X and Y.

Because these features are common to every member of the equivalence class of DAGs that could have faithfully generated a distribution for which P_O is the marginal, this feature of the causal structure can be tested, provided one is willing to assume faithfulness.

If one is not willing to assume faithfulness, then even this feature of the generating causal structure cannot be inferred. For example, if the DAG RS3 is interpreted as a linear structural equation model in which all variables are normally distributed with mean 0 and variance 1, then if a = -bc then $X \amalg Y$, i.e., the distribution is unfaithful and has produced an independence by a reduction in the dimensionality of the parameter space.

In general, the goal of causal inference procedures based on the work of Verma and Pearl and SGS is to output the features of the causal structure that are shared by every member of an equivalence class of causal structures, where every member of the class faithfully generates the independence relations found to hold in a sample of measured data So. SGS, and more recently Richardson (1996), have defined a graphical representation of d-separation equivalence classes, Partial Ancestral Graphs (PAGs), from which one can straightforwardly read off either features of the generating causal structure or the precise nature of our causal ignorance. In a PAG, each pair of variables X and Y are connected by one of the following kinds of edges, along with their causal interpretations:

```
X     Y X and Y are not directly causally connected
X o-o Y X and Y are causally connected (cc), but in an unknown way
X o-> Y X and Y are cc, and Y is not a cause of X
X <-o Y X and Y are cc, and Y is not a cause of X
X <-> Y X and Y are cc, X is not a cause of Y, and Y is not a cause of X
X ->  Y X is a cause of Y
```

The interpretations extend to every member of the equivalence class of DAGs that could have faithfully generated the independence relations in Po. The FCI algorithm (SGS, Scheines, et al., 1994) takes independence facts and outputs a PAG that represents the features of the generating structures that can be determined under the faithfulness assumption. So, for example, given $X \amalg Y$, the FCI algorithm would output the PAG: $X\ Y$, which means that X and Y are not causally connected. Given $\sim X \amalg Y$ then the FCI algorithm would output: $Xo{-}oY$, which means that X and Y are causally connected, but in an unknown way. Even though both of these PAGs represent an infinity of equivalent causal structures, they

each effectively choose one member of the d-separation partition of all the generating graphs that involve X and Y. Given three measured variables $\{X, Y, Z\}$ and two independence facts: $X \amalg Y$ and $Y \amalg Z$, then FCI would output:

```
X o-> Y <-o Z
```

which means that:

i) X and Z are not causally connected

ii) X and Y are causally connected, but Y is not a cause of X, and

iii) Y and Z are causally connected, but Y is not a cause of Z.

Thus, although we can infer nothing about the causes that do exist, by assuming faithfulness we can infer that Y is not a cause of either X or Z, a feature of the generating structure that in many circumstances would be of large practical value. Suppose that we assume that X is prior in time to Y, which is prior in time to Z, and we use this knowledge to exclude the possibility that Y is a cause of X and that Z is a cause of X or Y. Suppose further that from sample data we decide that $X \amalg Y|Z$. The the PAG output by the FCI algorithm is:

```
Xo-> Y -> Z
```

which means that:

i) X and Y are causally connected, and Y is not a cause of X, and

ii) X and Z are not directly causally connected, and

iii) Y is a cause of Z.

Although there might be an infinity of unobserved confounders influencing X and Y, in no member of the equivalence class is there any confounder between Y and Z, nor is there any member in which Z is a cause of Y. If there were, X and Y would not be d-separated by Z, and $X \amalg Y|Z$ would mean that the distribution was unfaithful. Without the assumption of faithfulness, from the same facts we could only conclude that:

i) X and Y are causally connected

ii) X and Z are causally connected

iii) Y and Z are causally connected

but we would have no further knowledge of the features of the generating structure. Given the independence facts true of the generating distribution, and the Markov and faithfulness assumptions, then the output of the FCI algorithm is an equivalence class that is guaranteed to include the generating DAG.

3. CONSISTENT TESTS

At the heart of these causal procedures is some test or model search technique for choosing between alternative causal models. In this section we review some basic definitions and results about statistical tests in general. We shall mainly be concerned with asymptotic tests and their relationship to finite sample sizes.

3.1. Consistent Tests. Suppose we have a model consisting of a set of probability measures $\Omega = \{P_\theta; \theta \in \Omega\}$ where each P_θ lives on a sample space $\mathcal{V}$. We observe i.i.d. data $V^n = (V_1, \ldots, V_n)$ from some probability measure P in the model. Let $\mathcal{V}^n = \mathcal{V} \times \cdots \times \mathcal{V}$. For brevity, we write P instead of P^n for the product measure.

The "null hypothesis" is a subset $\Omega_0 \subset \Omega$. Suppose we want to test:

$$H_0 : P \in \Omega_0 \quad \text{versus} \quad H_1 : P \notin \Omega_0.$$

As usual, a test consists of a rejection region $R_n \subset \mathcal{V}^n$; if $V^n \in R_n$ we reject H_0. We shall be studying the asymptotic properties of tests. Thus, suppose we specify a test R_n for each sample size n and let $R = (R_1, R_2, \ldots)$.

Sometimes we do not observe V but rather we observe some function of $Y = t(V)$. For example, $V = (\mathcal{O}, \mathcal{U})$ and we only observe $\mathcal{O}$. Obviously, the test can be based only on the observable $Y = t(V)$. In these cases, it will be understood that the rejection regions R_n are subsets of $\mathcal{Y}^n = \mathcal{Y} \times \cdots \times \mathcal{Y}$ or, in other words, that the indicator function for R_n is Y^n measurable. In what follows, all limits refer to the sample size n tending to ∞. When we refer to a test, we mean the sequence of tests R.

DEFINITION 1. A test is **pointwise consistent** if
(i) for every $P \in \Omega_0$, $P(R_n) \to 0$ and
(ii) for every $P \notin \Omega_0$, $P(R_n) \to 1$.

To discuss the notion of uniform consistency, we need to introduce a topology on Ω. For this we use the total variation metric defined by $d(P, Q) = \sup_A |P(A) - Q(A)|$. Also define $d(P, \Omega_0) = \inf_{Q \in \Omega_0} d(P, Q)$ and for every $\delta > 0$ let $\Omega_\delta = \{P; d(P, \Omega_0) \leq \delta\}$.

DEFINITION 2. A test is **uniformly consistent** if

(i) $\sup_{P\in\Omega_0} P(R_n) \to 0$ and
(ii) for every $\delta > 0$, $\sup_{P\in\Omega_\delta^c} P(R_n) \to 1$.

Remark: To avoid triviality, if Ω_0 is dense in Ω then we say that no uniformly consistent test exists.

The difference between a pointwise and uniformly consistent test is the presence of the suprema in the definitions. The reason why the difference matters is discussed in Section 3.4.

3.2. Hypotheses About Causal Graphs. The definitions above do not immediately relate to hypotheses about causal graphs. We make this connection as follows. Suppose that G is a DAG. A sub-graph is a DAG obtained from G by deleting one or more arrows in G. Let A and B be two sets of sub-graphs. Let Ω_A be all distributions faithful to some DAG in A and let Ω_B be all distributions faithful to some DAG in B. Finally, let $\Omega = \Omega_A \cup \Omega_B$. If $\Omega_A \cap \Omega_B = \emptyset$ then we say that (A, B) is a proper dichotomization of G.

In the remainder of the paper, we shall consider hypotheses about sets of subgraphs. When we do so, it is understood that these can be translated into statements about sets of distributions, as described in the previous paragraph.

3.3. Importance of Uniform Consistency. Pointwise and uniformly consistent tests both guarantee good behavior with large samples, but there are important practical differences between these modes of convergence.

First, uniform consistency is what links asymptotic (i.e. large sample) procedures to finite samples. To see this, suppose that, given $\epsilon > 0$, we wish to find a sample size $n_0(\epsilon)$ such that the probability of falsely rejecting the null hypothesis is bounded above by ϵ if $n \geq n_0(\epsilon)$. To achieve this goal with a test that is pointwise but not uniformly consistent requires one to know the true data generating probability. That is, $n_0(\epsilon)$ is a function of the unknown P. Put another way, if the test is pointwise but not uniformly consistent, then for any ϵ,

$$\inf\{n_0; P^n(R_n) \leq \epsilon, \text{for all } P \in \Omega_0 \quad \text{and} \quad \text{for all } n \geq n_0\} = \infty.$$

There is no finite sample size that bounds the probability of an error. If a test if uniformly consistent then we can find $n_0(\epsilon)$ which does not depend on P and such that $\sup_{P\in\Omega_0} P^n(R_n) \leq \epsilon$ for all $n \geq n_0(\epsilon)$. Furthermore, if the test is uniformly consistent, then given an error rate

$\epsilon > 0$ and a $\delta > 0$, we find an n_0 such that, for all $n \geq n_0$, the probability of falsely rejecting the null is bounded above by ϵ and the probability of failing to reject when $d(P, \Omega_0) > \delta$ is also bounded above by ϵ for all $n \geq n_0$, without knowledge of P.

Second, there is a relationship between tests and confidence intervals. Loosely speaking, tests cannot be inverted to form confidence intervals unless they are uniformly consistent. We discuss this further in Section 5.

In the next section we show that it is possible to construct uniformly consistent and strongly consistent tests of associations but that it is difficult to do so for causal effects.

4. NON-EXISTENCE OF UNIFORMLY CONSISTENT TESTS FOR CAUSAL HYPOTHESES.

In this section we prove the non-existence of uniformly consistent tests for causal hypotheses in the two paradigmatic examples from Section 2. Recall that the first case involves a potential cause X, an outcome Y and a potential confounder U. The second case involves a covariate X, a potential cause Y, an outcome Z and two potential confounder U and V.

4.1. Case 1: Two Variables. The starting point here are two observed random variables X and Y. Recall that from observing X and Y to be independent in a large sample, the conclusion from the DAG framework is that X does not cause Y. Suppose that $V = (X, Y, U)$, that X and Y are known to be time ordered and that X precedes Y. The variable U is meant to represent possible confounding variables. There are in fact infinitely many choices V in which to embed X, Y. But one latent variable U is enough to demonstrate the lack of a uniformly consistent test.

The question of interest is whether X causes Y i.e. is there an arrow from X to Y? To be concrete, we will take X and Y to be binary and we take U to be discrete. We assume that the random variable U takes at least four distinct values. Let G be the complete DAG for V (with an arrow from X to Y) as in Figure 1.

Now we have to define what hypothesis we shall test. Pursuant to the discussion in Section 3, this means we need to dichotomize the sub-graphs into two disjoint sets of subgraphs A and B, say. Since we want to know whether X causes Y it seems natural to take A to be all distributions faithful to any sub-graph with an arrow from X to Y and to take B to be the complement. However, contrary to intuition, this is not a fruitful way to proceed, and it is not the way that procedures like those in SGS work. The reason why this natural dichotomy

is not appropriate is deferred until section 4.4. Instead, we dichotomize based on whether or not $X \amalg Y$ holds or not. Specifically, let A be all subgraphs with an arrow from X to Y and also let A include the sub-graph where there is no arrow from X to Y but there is an arrow from U to X and from U to Y. Let B be all other subgraphs. See Figure 2. Assuming faithfulness, the set of distributions in B is equivalent to $X \amalg Y$ and the set of distributions in A is equivalent to "not $X \amalg Y$".

Note that if we accept the hypothesis B then we can conclude that "X does not cause Y" since each graph in B has no arrow from X to Y. If we accept A our conclusion is "we cannot make a decision about whether X causes Y" since some graphs in A have arrows from X to Y while others do not. Thus, the best we can do is discover non-causation. In Section 4.2 we consider a case where there is the possibility of discovering causation.

The results about the existence of pointwise and uniformly consistent tests depend on whether or not U is observed and on whether A or B is treated as the null hypothesis. In practice, we are really interested in the case where U is unobserved. We include the case where U is observed to make it clear exactly what is lost by the presence of unobserved confounding. The proof of Theorem 1 and all other results are deferred until section 8.

THEOREM 1. *The existence or non-existence of pointwise consistent and uniformly consistent tests is as summarized in Table 1.*

H_0	U observed?	Pointwise Consistent?	Uniformly Consistent?
$P \in \Omega_A$	$\surd$	$\surd$	$\times$
$P \in \Omega_B$	$\surd$	$\surd$	$\surd$
$P \in \Omega_A$	$\times$	$\surd$	$\times$
$P \in \Omega_B$	$\times$	$\surd$	$\times$

Table 1. Summary of results for Theorem 1. $\surd$ = Yes and $\times$ = No.

The cases where $H_0 : P \in \Omega_A$ do not have uniformly consistent tests since A^c is dense in Ω. The more interesting case is when $H_0 : P \in \Omega_B$. In this case, if U we observed, we could end up concluding that X is not a cause of Y using a uniformly consistent test. We cannot do so of U is unobserved as in an observational study.

It is worth recalling that there do exist uniformly consistent tests for testing associations.

PROPOSITION 1. *Let Ω_0 be all P's such that $X \amalg Y$ under P and let Ω_1 all other distributions in $\mathcal{P}$. Then there exists a uniformly consistent test for $H_0 : P \in \Omega_0$ versus*

$H_1 : P \in \Omega_1$.

4.2. Case 2. Now consider time ordered variables X, Y and Z and potential confounding variables U and V. We are interested in the causal effect of Y on Z. The directed acyclic graph to describe the model is given in Figure 3. Again, we take X, Y and Z be to be binary and we take U and V to be discrete, each taking at least four unique values.

The dichotomization is as follows. Let B consist of all subgraphs in which (i) there is an arrow from X to Y, (ii) there is an arrow from Y to Z, (iii) there is at most one arrow emanating from U and (iv) there is at most one arrow emanating from V. Let A be all other subgraphs. If we accept the hypothesis B then we can conclude that Y causes Z. If we accept A then the conclusion is "no decision about causation." In this sense, this case is the dual of Case 1. Some further intuition on this dichotomy is given in the next Lemma.

LEMMA 1. *Let Ω_A be all distributions faithful to some graph in A and let Ω_B be all distributions faithful to some graph in B. Then, assuming faithfulness, $P \in B$ if and only if, under P we have*

$$(X \amalg Z|Y) \text{ and } (\text{not } X \amalg Y) \text{ and } (\text{not } Y \amalg Z).$$

THEOREM 2. *Under the above dichotomization, the existence or non-existence of pointwise consistent and uniformly consistent tests is as summarized in Table 2.*

H_0	U observed?	Pointwise Consistent?	Uniformly Consistent?
A	√	√	×
B	√	√	√
A	×	√	×
B	×	√	×

Table 2. Summary of results for Theorem 2. √ = Yes and × = No.

4.3. Why Randomized Experiments Do Have Uniformly Consistent Tests. In a randomized experiment, it is possible to construct uniformly consistent tests of causal hypotheses. To see this, return to case 1 of Section 4.1. Randomization breaks the arrow

from U to X in Figure 1. The hypothesis H_0 : "no arrow from X to Y" is then equivalent to $H_0 : X \amalg Y$ and there are well known uniformly consistent tests, as in Proposition 1.

4.4. Other Choices of Dichotomizations. The choice of dichotomizations in Sections 4.1 and 4.2 might seem odd. For example, in case 1, it might seem more natural to take A to be all graphs with an arrow from X to Y and B to be all graphs with no arrow from X to Y. Then A corresponds to "X causes Y" and B corresponds to "X does not cause Y". Let us call this the "natural dichotomization." However, it is easy to see that this choice makes the situation worse. Indeed, there will no longer exist pointwise consistent tests. The reason is due to the graph G_* in which there is no arrow from X to Y but there are arrows from U to X and from U to Y. In the natural dichotomization, G_* is moved from A to B. But the (X, Y) marginal distribution under G_* is indistinguishable from that of the graph $X \to Y$ in A. This makes the two hypotheses indistinguishable. We state this formally below.

THEOREM 3. *Using the natural dichotomy described above, there are no pointwise or uniformly consistent tests if U is unobserved.*

In fact, it can be shown that the dichotomy in Section 4.1, which is the dichotomy used by SGS, is the only one that leads to a non-trivial conclusion and has pointwise consistency. Similar comments apply to Case 2.

5. CONFIDENCE INTERVALS AND POINT ESTIMATES

So far we have confined attention to testing. In practice, we are often more interested in point estimation and interval estimation. To proceed, then, we need to define what exactly we are estimating.

The parameter of interest in causal problems is the causal effect or treatment effect. It is defined precisely in Robins (xxxx) using a functional called the G-equation and is also defined in SGS using the manipulation theorem. Both approaches lead to the same formula.

In case 1, (one version of) the causal effect is given by

$$\theta = \int [Pr(Y = 1|X = 1, U = u) - Pr(Y = 1|X = 0, U = u)] dP_U(u)$$

which can be thought of as the proportion of the population who would have $Y = 1$ if X were set to 1, minus the proportion of the population who would have $Y = 1$ if X were set to 0. In the language of counterfactuals (Neyman xxxx, Rubin xxxx, Robins xxxx), this is equivalent to $\theta = E(Y_1) - E(Y_0)$ where Y_1 is the outcome of a subject when $X = 1$ and Y_0

is the outcome of a subject when $X = 0$. When we want to be clear that θ depends on P we write $\theta = T(P)$.

In case 2 the causal effect is

$$\theta = \int [Pr(Z = 1|Y = 1, U = u, V = v) - Pr(Z = 1|Y = 0, U = u, V = v)] dP_{U,V}(u, v).$$

In each case, note that $\theta \in \Theta = [-1, 1]$.

5.1. Confidence Intervals. An asymptotic $1 - \alpha$ confidence region I_n is a function of the observable into the subsets of Θ such that

$$\text{liminf}_n \inf_{P \in \Omega} P(T(P) \in I_n) \geq 1 - \alpha.$$

A confidence region is consistent if it eventually omits all false values i.e. if $T(Q) \neq T(P)$ implies that

$$P(T(Q) \in I_n) \to 0.$$

THEOREM 4. *Let $\alpha \in (0, 1)$. In cases 1 and 2 from Section 4, if U is not observed then there do not exist consistent $1 - \alpha$ confidence regions for θ.*

5.2. Point Estimation. A point estimate is a function $\hat{\theta}_n$ of the observable data into Θ. A point estimate is consistent if, for every P and every $\epsilon > 0$,

$$P(|T(P) - \hat{\theta}_n| > \epsilon) \to 0.$$

A point estimate is uniformly consistent if, for every compact set $K \subset \Omega$, and for every $\epsilon > 0$,

$$\sup_{P \in K} P(|T(P) - \hat{\theta}_n| > \epsilon) \to 0.$$

THEOREM 5. *Consider the two cases studied in Section 4. In both cases, there does not exist a uniformly consistent point estimate.*

8. PROOFS

Proof of Theorem 1. Pointwise Consistency Results.

Case 1: $H_0 = B$; U is observed. Let Λ_n be the standard likelihood ratio statistic based on (X^n, Y^n) for testing $\tilde{H}_0 : A \amalg B$ versus $\tilde{H}_1 :$ not $A \amalg B$. Clearly, a pointwise consistent test for $\tilde{H}_0$ versus $\tilde{H}_1$ is also a pointwise consistent test for H_0 versus H_1. Let

$R_n = \{(X^n, Y^n); \Lambda_n \geq c_n(\alpha_n)\}$ where $c_n(t)$ is such that $Pr(\chi_1^2 > c_n(t)) = t$, and $\alpha_n \to 0$. Standard asymptotic theory shows that, under $\tilde{H}_0$, Λ_n converges in distribution to a χ_1^2 random variable. Hence, for any $P \in \Omega_0$, $P^n(R_n) = \alpha_n + o(1) \to 0$.

On the other hand, if $P \in \tilde{H}_1$ then again a standard result from asymptotic theory of likelihood ratio tests shows that the test has asymptotic power 1, i.e. $P^n(R_n) \to 1$.

Case 2: $H_0 = A$; U is observed. Use the same argument but define the rejection region to be the complement of R_n as defined above.

Case 3: $H_0 = B$; U is unobserved. Note that the proof for the case where U was observed did not make use of U, i.e. Λ_n is (X^n, Y^n) measurable. So the same proof applies.

Case 4: $H_0 = A$; U is unobserved. Same as above.

Proof of Theorem 1: Uniform Consistency Results.

Case 1: $H_0 = B$; U is observed. Note that $P \in B$ if and only if

$$X \amalg Y \quad \text{and} \quad X \amalg Y|U.$$

Let Λ_n be the likelihood ratio test for $X \amalg Y$ and let Γ_n be the likelihood ratio test for $X \amalg Y|U$. (The latter is a function of (X^n, Y^n, U^n) which is permitted since U is observed.) Standard asymptotic theory shows that a uniformly consistent test for $X \amalg Y$ can be constructed by taking $R_n = \{(X^n, Y^n, U^n); \Delta_n \geq c_n\}$ for an appropriate choice of c_n. Similarly, a uniformly consistent test for $X \amalg Y|U$ can be constructed by taking $S_n = \{(X^n, Y^n, U^n); \Gamma_n \geq d_n\}$ for an appropriate choice of c_n. Let $T_n = R_n \cup S_n$. It is easy to see that T_n forms a uniformly consistent test for H_0.

Case 2: $H_0 = A$; U is observed. Since A^c is dense in Ω, the result follows from the remark after Definition 3.

Case 3: $H_0 = B$**;** U **is unobserved.** In what follows, if P is any distribution for (X, Y, U) then $\tilde{P}$ will denote the (X, Y) marginal of P. We will now find two distributions P and Q and a positive number $\delta > 0$ such that (i) $d(P, \Omega_0) > \delta$, (ii) $Q \in \Omega_0$, (iii) $\tilde{P} = \tilde{Q}$, (iv) P is unfaithful, (v) for all small $\epsilon > 0$, there exists a faithful $P_1 \in \Omega_1$ such that $d(P, P_1) < \epsilon$.

In Table 3 we list the 16 atoms of the sample space and the mass function p of P.

X	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
Y	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
U	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
p	0	0	0	1/4	0	0	1/4	0	0	1/4	0	0	1/4	0	0	0
r	sqa_1	sqa_2	sqa_3	sqa_4	$s\overline{q}a_1$	$s\overline{q}a_2$	$s\overline{q}a_3$	$s\overline{q}a_4$	$\overline{s}qa_1$	$\overline{s}qa_2$	$\overline{s}qa_3$	$\overline{s}qa_4$	$\overline{sq}a_1$	$\overline{sq}a_2$	$\overline{sq}a_3$	$\overline{sq}a_4$

Table 3. Mass Functions for Proof of Theorem 1.

Let Q be the uniform distribution. Then $Q \in \Omega_0$ and it is easy to see that $\tilde{P} = \tilde{Q}$. Thus, (ii) and (iii) hold. Also, direct calculations show that, under P, X is not independent of U and Y is not independent of (X, U). On the other hand, from (iii) and the fact that $X \amalg Y$ under Q, we see that $X \amalg Y$ under P. Thus, P is unfaithful so (iv) holds.

Next we verify (i). Let $\delta = 1/32$. We can decompose Ω_0 into three disjoint sets Ω_{01}, Ω_{02} and Ω_{03} where $R \in \Omega_{01}$ iff (X, Y, U) are mutually independent under R, $R \in \Omega_{02}$ iff X and U are dependent and Y is independent of (X, U) and, $R \in \Omega_{03}$ iff Y and U are dependent and X is independent of (Y, U). Consider a R in Ω_{01}. Then R must have a mass function r of the form shown in Table 3, where $s, q \in [0, 1]$, $a_1, a_2, a_3, a_4 \geq 0$, $\sum_{j=1}^{r} a_j = 1$, $\overline{s} = 1 - s$ and $\overline{q} = 1 - q$. A well known property of total variation distance is that $d(P, Q) = \frac{1}{2} \sum_z |p(z) - q(z)|$ the sum being over all points in he sample space. Thus,

$$\begin{aligned} 2d(P, R) &= sq(1 - a_4) + s\overline{q}(1 - a_3) + \overline{s}q(1 - a_2) + \overline{sq}(1 - a_1) \\ &+ \left|\frac{1}{4} - sqa_4\right| + \left|\frac{1}{4} - s\overline{q}a_3\right| + \left|\frac{1}{4} - \overline{s}qa_2\right| + \left|\frac{1}{4} - \overline{sq}a_1\right|. \end{aligned}$$

Suppose first that $p \geq 1/2$ and $s \geq 1/2$. If $a_3 sq \geq \delta$ then $2d(P, Q) \geq \delta$. If not, then $a_3 < \delta/(sq) \leq 4\delta$. Hence,

$$s\overline{q}a_3 \leq \frac{s}{2} a_3 \leq 2\delta = \frac{1}{16}.$$

Thus,

$$\left|\frac{1}{4} - s\overline{q}a_3\right| \geq \frac{1}{4} - \frac{1}{16} > \delta.$$

In either case then, $2d(P, Q) \geq \delta$. When one or both of s and q is not greater than or equal to $1/2$, a similar argument shows that $2d(P, Q) \geq \delta$. We have thus demonstrated that

$d(P, \Omega_{01}) \geq 1/64$. It can be shown that $d(P, \Omega_{02}) \geq 1/64$ and $d(P, \Omega_{03}) \geq 1/64$ using similar arguments. The details are omitted.

Fix $\epsilon > 0$. It is possible to find a faithful distribution P_1 such that $d(P, P_1) < \epsilon$. This can be seen by perturbing the mass function of P by a small amount and using continuity of $d(P, R)$. Note that $d(P_1, \Omega_0) > \delta - \epsilon$ so that $P_1 \in \Omega^c_{\delta/2}$ for ϵ small.

Suppose there exists a uniformly consistent test with rejection regions $R_1, R_2, \ldots$. Note that since U is unobserved, the test is (X^n, Y^n) measurable. Now,

$$\begin{array}{rcll} Q(R_n) & = & \tilde{Q}(R_n) & \text{since } R_n \text{ depends only on } (X^n, Y^n) \\ & = & \tilde{P}(R_n) & \text{since } \tilde{P} = \tilde{Q} \\ & = & P(R_n) & \text{since } R_n \text{ depends only on } (X^n, Y^n) \\ & \geq & P_1(R_n) - \epsilon & \text{since } d(P, P_1) \leq \epsilon. \end{array}$$

Since the test is consistent and since $P_1 \in \Omega^c_{\delta/2}$, $P_1(R_n) \to 1$. Hence, $\liminf_n Q(R_n) \geq 1 - \epsilon$. Since this holds for all $\epsilon > 0$, it follows that $\lim_n Q(R_n) = 1$. Thus, $\lim_n \sup_{P \in \Omega} P(R_n) \geq \lim_n Q(R_n) = 1$ contradicting the fact that $\lim_n \sup_{P \in \Omega} P(R_n)$ should tend to 0 for a uniformly consistent test.

Case 4: $H_0 = A$; U is unobserved. Since A^c is dense in Ω, the result follows from the remark after Definition 3.

This completes the proof of Theorem 1. □

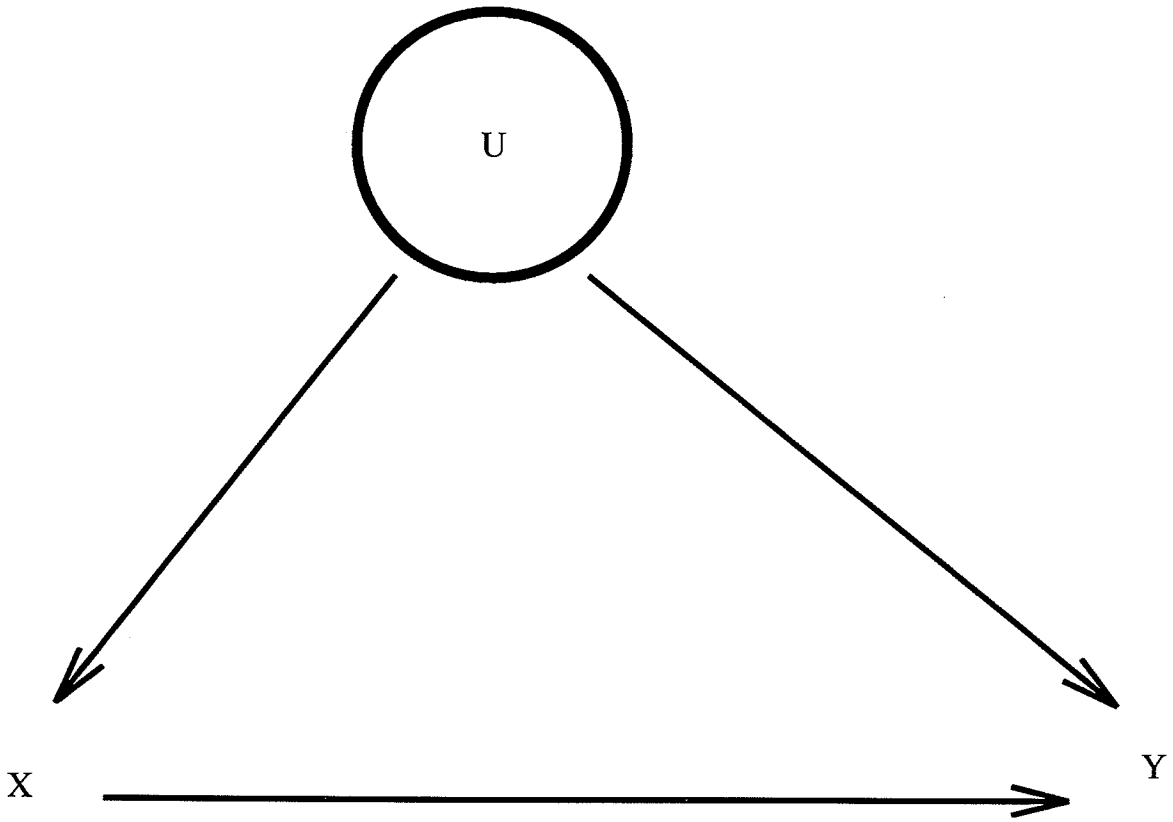

Figure 1. Directed Acyclic Graph for Case 1.

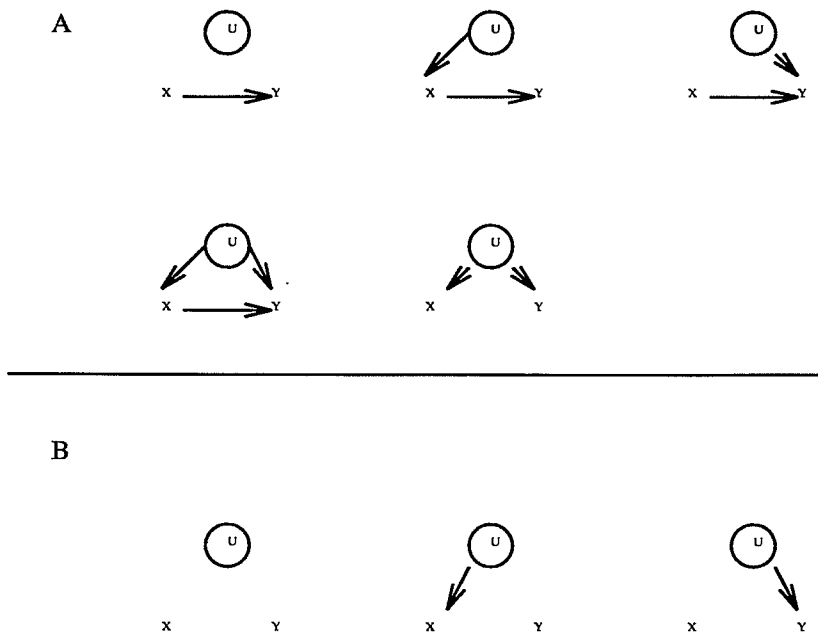

Figure 2. Dichotomization for Case 1.

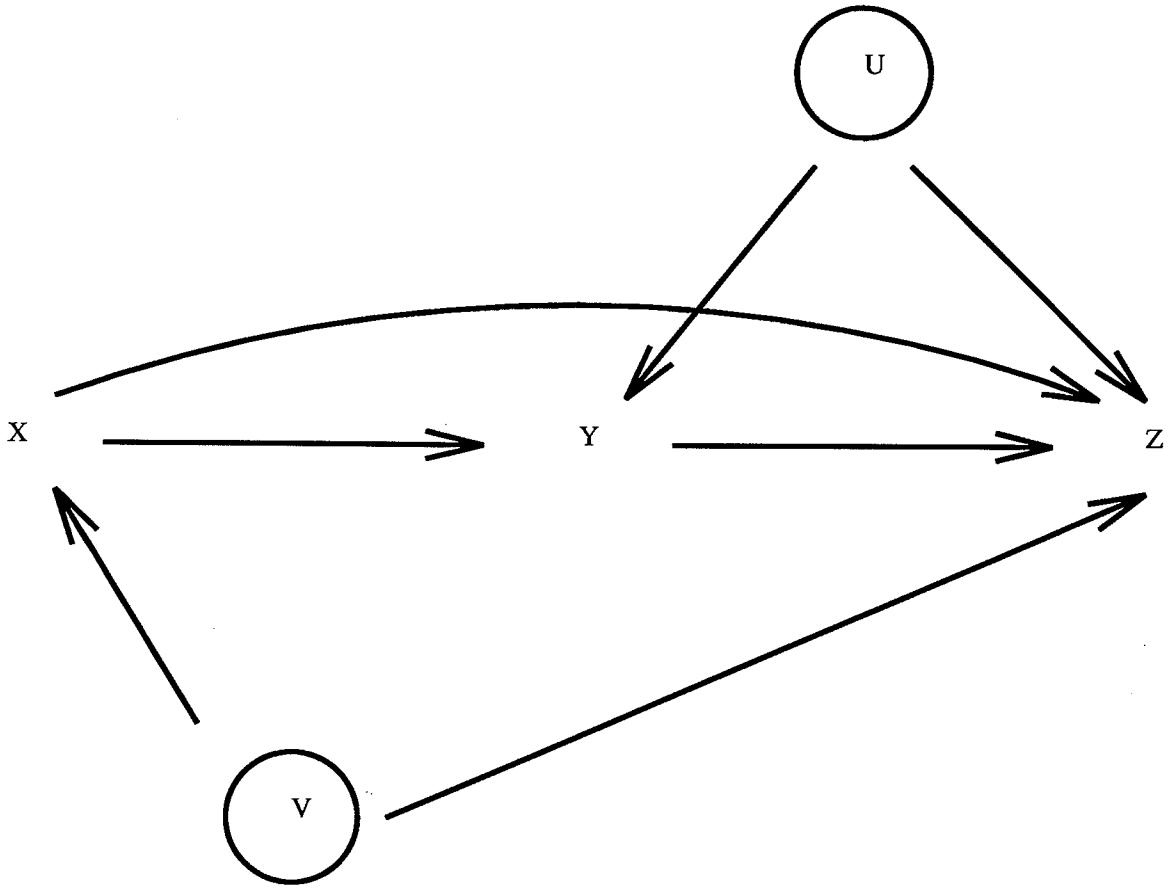

Figure 3. DAG for Case 2.

REFERENCES

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669-709.

Pearl, J. and Verma, T. (1991). A theory of inferred causation. In *Principles of Knowledge, Representation and Reasoning: Proceedings of the Second International Conference.* (J.A. Allen, R. Filkes and E. Sandewall, eds.) p. 441-452.

Robins, J.M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, **7**, 1393-1512.

Robins, J.M. (1987). Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods – Application to control of the healthy worker survivor effect." *Computers and Mathematics with Applications*, 14:923-945.

Robins, J.M. (1995b). "Discussion of 'Causal Diagrams for empirical research' by J. Pearl," *Biometrika*, 82, 695-698.

Robins, J.M. (1996). Causal inference from complex longitudinal data. *Proceedings of the UCLA Conference on Causal Modelling in Latent Variables*, to appear.

Rubin, D. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688-701.

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