Uniform Consistency In Causal Inference

James M. Robins, Richard Scheines, Peter Spirtes and Larry Wasserman[1]
Harvard University and Carnegie Mellon University

September 1, 2000

There is a long tradition of representing causal relationships by directed acyclic graphs (Wright 1934). Spirtes (1992) and Spirtes, Glymour and Scheines (1993) using some ideas in Pearl and Verma (1991) describe procedures for inferring the presence or absence of causal arrows in the graph even if there might be unobserved confounding variables, and/or an unknown time order, and that under weak conditions, for certain combinations of DAGs and probability distributions, are asymptotically (in sample size) consistent. These results are surprising since they seem to contradict the standard statistical wisdom that consistent estimates of causal effects do not exist for non-randomized studies if there are potentially unobserved confounding variables.

We resolve the apparent incompatibility of these views by closely examining the asymptotic properties of these causal inference procedures. We show that the asymptotically consistent procedures are "pointwise consistent" but "uniformly consistent" tests do not exist. Thus, no finite sample size can ever be guaranteed to approximate the asymptotic results. We also show the non-existence of valid, consistent confidence intervals for causal effects and the non-existence of uniformly consistent point estimates. Our results make no assumptions about the form of the tests or estimates. In particular, the tests could be classical independence tests, they could be Bayes tests or they could be tests based on scoring methods such as BIC or AIC. The implications of our results for observational studies are controversial and are discussed briefly in the last section of the paper.

The results hinge on the following fact: it is possible to find, for each sample size n, distributions P and Q such that: P and Q are empirically indistinguishable and yet P and Q correspond to different causal effects.

1. INTRODUCTION

The problem of inferring causal relationships between variables is a topic of continuing interest. Common statistical wisdom dictates that causal effects cannot be consistently estimated from observational studies alone unless one observes and adjusts for all possible confounding variables, and knows the time order in which events occurred. However, Spirtes (1992) and Spirtes, Glymour and Scheines (1993) – using some ideas in Pearl and Verma (1991) – developed a framework in which causal relationships are represented by arrows in a directed acyclic graph G. They also described asymptotically consistent procedures for determining features of causal structure from data even if we allow for the possibility of unobserved confounding variables and/or an unknown time order. For certain combinations of DAGs and probability distributions, these procedures can infer the existence or absence of causal relationships. In particular, Spirtes, Glymour, and Scheines (1993)

[1]Spirtes' research was supported by NSF grant DMS-9873442. Wasserman's research was supported by N.I.H. grant R01-CA54852-01 and NSF grants DMS-9303557 and DMS-9357646.

proved the Fisher consistency of these procedures. Pointwise consistency follows from the Fisher consistency and the uniform consistency of the test procedures for conditional independence relation that the procedures use.

We will call the framework assumed in Spirtes, Glymour and Scheines (1993) the SGS causal model. The SGS model is closely related to Wright's path diagrams (Wright 1934) and structural equation models with acyclic path diagrams and no correlated errors (Bollen, 1989, Spirtes et al., 1993, Pearl, 1995). The SGS model has also been extended to allow cyclic directed graphs and correlated errors, although this is not described here. See Spirtes et al. (1998) and Richardson (1996). The reader is referred to Lauritzen (1996) and Wermuth (1980) for further details on statistical graphical models. A related approach to causation is based on the theory of counterfactuals; for a recent discussion on this approach, see Dawid (2000).

This paper explores the apparent discrepancy between the common statistical wisdom and the aforementioned results. We show that under the SGS model there do, in certain canonical cases, exist "pointwise consistent" tests but there do not exist "uniformly consistent" tests for causal effects. This implies that these tests are guaranteed to yield correct answers with an infinite sample size, but that no test can make such guarantees in finite samples, even approximately, no matter how large the sample. Further, we show that there are no valid, consistent confidence intervals for causal effect parameters in the presence of potential confounding variables. The methodological implications of these results for observational studies are controversial, and are briefly discussed in section 9.

In Section 2 we review the SGS model. In Section 3 we introduce several canonical examples. In Section 4 we discuss consistent tests. In Section 5 we present the main result of the paper: the non-existence of uniformly consistent causal inference procedures. Section 6 examines the implications for confidence intervals and point estimates. Section 7 discusses Bayes' tests. In Section 8 we discuss the problem of inferring time order. Some further remarks are in Section 9. Proofs of results are in Section 10. To make the results precise, it is necessary to introduce a certain amount of notation. For the reader's convenience, we have included a glossary of notation in the Appendix. The appendix also defines the notion of "d-separation". Throughout the paper, we write $X \amalg Y$ to mean that the random variables X and Y are independent. We write $X \amalg Y|Z$ to mean that the random variables X and Y are independent given Z (where X, Y, and Z may represent individual random variables or sets of random variables.)

2. THE SGS MODEL FOR CAUSAL INFERENCE

A DAG G is a set of vertices with arrows between some pairs of vertices such that there are no directed cycles, i.e. one cannot start at a vertex and follow a directed path back to that vertex. (A directed path between vertices A and B is a sequence of vertices starting at A and ending at B of the form $A \to X_1 \to \cdots \to X_k \to B$ or of the form $A \leftarrow X_1 \leftarrow \cdots \leftarrow X_k \leftarrow B$.) An example of a DAG is given in Figure 1.

The SGS model begins with a triple (G, V, P) where G is a DAG with a set of vertices V, $V = (X_1, \ldots, X_k)$ is a vector of random variables and P is a joint probability distribution for V. The random variable X_j takes values in a set $\mathcal{X}_j$. Arrows represent causal relationships: if there is an arrow pointing from X_i to X_j it means that X_i has a direct causal effect on X_j (relative to V). In what follows we will always assume that the distribution P is "Markov" to G and we will sometimes assume that P is "faithful" to G. We now proceed to define these two concepts.

X_i is in the set of **parents** of a variable X_j in a DAG G, denoted by $PA_G(X_j)$, if there is an edge $X_i \to X_j$ in G. X_j is a **descendent** of X_i in G if there is a directed path from X_i to X_j or $X_i = X_j$. A distribution P with density function p is **Markov** to G if

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

where the latter formula is called the **Markov factorization** of P according to G. Equivalently, P is Markov to G if for every subset W of the variables V, we have that $X_i \amalg \hat{X}_i | PA_G(X_i)$ where $\hat{X}_i$ are all variables in V that are not descendents or parents of X_i in G. If each of the conditional probabilities $p(x_i | pa_G(x_i))$ in the Markov factorization is well defined, then P has **full Markov support** relative to G. Let $\mathcal{P}(G)$ denote all distributions that are Markov to G. If P is in $\mathcal{P}(G)$ we say that P is **compatible** with G. The Markov assumption is that if G represents the data generating mechanism for P, then $P \in \mathcal{P}(G)$.

Given $P \in \mathcal{P}(G)$, let $\mathcal{I}(P)$ represent all independence and conditional independence relations that hold for the variables in V under P. Let $\mathcal{I}_G = \cap_{Q \in \mathcal{P}(G)} \mathcal{I}(Q)$ be all independence relations that are common to all the distributions in $\mathcal{P}(G)$. We say that P is **faithful** to G if $\mathcal{I}(P) = \mathcal{I}_G$. Otherwise, P is called **unfaithful**. In other words, P is faithful if it does not possess "extra" independence relations not shared by all the others distributions in $\mathcal{P}(G)$. In cases where $\mathcal{P}(G)$ can be parameterized by a family of distributions with a parameter of finite dimension, the set of unfaithful distributions typically has Lebesgue measure 0.

Let $\Omega(G) \subset \mathcal{P}(G)$ denote the set of all distributions that are faithful to G. The faithfulness assumption is that if G represents the data generating mechanism for P, then $P \in \Omega(G)$. As SGS say: "... the Faithfulness Condition can be thought of as the assumption that conditional independence relations are due to causal structure rather [than] to accidents of parameter values."

Let $\mathcal{G}$ be some set of directed acyclic graphs of interest. Define $\mathcal{P}(\mathcal{G}) = \cup_{G \in \mathcal{G}} \mathcal{P}(G)$ and $\Omega(\mathcal{G}) = \cup_{G \in \mathcal{G}} \Omega(G)$. In many of our examples, $\mathcal{G}$ has the following special structure. We start with time ordered random variables and we let G be the unique, complete graph consistent with this time ordering; complete means no missing arrows. Then $\mathcal{G}$ is G together with all sub-graphs of G where a sub-graph is obtained by deleting some of the arrows in G. This is the structure of $\mathcal{G}$ throughout the paper except in Section 8.

Typically, we do not observe all the random variables V. In this case we write $V = (O, U)$ where O are the observed random variables and U are the unobserved, or latent, random variables. Let $\widetilde{P}$ denote the marginal distribution of P for the observed O. We suppose we have i.i.d. observations $O^n = (O_1, \ldots, O_n)$ from $\widetilde{P}$.

We can now summarize the SGS model. We start with a set of graphs $\mathcal{G}$ and we assume there exists a triple (G, V, P) where $G \in \mathcal{G}$, $V = (O, U)$ and that P is Markov and Faithful to G. We observe n i.i.d. observation $O^n = (O_1, \ldots, O_n)$ from $\widetilde{P}$, the O-marginal of P. We shall address the following question: given two variables of interest X_i and X_j in O, and given a sample O^n from $\widetilde{P}$, can we test whether there is an arrow from X_i to X_j in G (or more precisely, is there a directed path from X_i to X_j in G which except for the endpoints contains only unobserved variables)? We also address a related question: can we estimate the size of the causal effect (as defined in Section 4.2) of X_i on X_j? It is critical to recognize that in general the analyst does not know the graph G that generated the data and the only available information are the data O^n, and possible background knowledge (such as time order.)

Even given the Markov and Faithfulness Assumptions, many distributions over the random variables in a DAG are compatible with more than one DAG which yield different predictions about the causal effect of a variable. For example, if there are just two measured variables X and Y which are dependent, the joint distribution over X and Y is compatible with X causing Y, or Y causing X. In the former case, the causal effect of X on Y is not zero, while in the latter case it is zero. When faced with this situation, the SGS procedures output "No conclusion." However, there are some distributions that are compatible only with a set of DAGs all of which yield the same prediction about the causal effect of a variable. The SGS approach is to characterize which distributions have this property, under a variety of background assumptions.

3. SOME EXAMPLES

Before proceeding with any technical details, we introduce some examples. In this section, we deal with the examples theoretically. Specifically, we assume that rather than given a sample O^n of size n, we are given the marginal distribution of O. In Sections 4 to 8, we examine the realistic setting in which the marginal distribution of O is unknown but we obtain a sample O^n.

Example 0. Let $V = (Z, X, Y)$ where X is number of cigarettes smoked in one year, Y is a measure of disease at a later time, and Z represents all potential common causes of X and Y. Suppose, for the purposes of this example, that the following structural equation model holds, where X, Y, and Z are standardized Normal variables:

$$Z = \epsilon_Z$$
$$X = \alpha Z + \epsilon_X$$
$$Y = \beta X + \gamma Z + \epsilon_Y$$

ϵ_X, ϵ_Y, ϵ_Z are Normally distributed with mean 0. The DAG G is given in Figure 1. In this model, X is a direct cause of Y if and only if $\beta \neq 0$. In other words, there is an arrow from X to Y if and only if $\beta \neq 0$. Suppose we do not observe Z. Thus, $O = (X, Y)$ and $U = Z$. Suppose we are given that X and Y are uncorrelated. We will show that this information is sufficient to deduce that X is not a cause of Y under the SGS model. First note that there are two explanations for the zero correlation. One possibility is that β (and at least one of α and γ) is 0. In this case, inferring that X does not cause Y from the data is correct. On the other hand, even if β is large we can still get a zero correlation between X and Y. This will happen if $\beta = -\alpha\gamma$. For example, we might have $\beta = 0.72$, $\alpha = 0.9$ and $\gamma = -0.8$. In this case, X does cause Y – indeed, the causal effect β is large (compared to the variance) – and inferring that X does not cause Y would be an error. Intuitively, it seems "unlikely" that $\beta = -\alpha\gamma$. Specifically, the set of (α, β, γ) values that satisfy $\beta = -\alpha\gamma$ have Lebesgue measure 0 in R^3. A distribution for which $\beta = -\alpha\gamma$ is unfaithful to G since it has an independence between some variables not because of missing arrows in the DAG but because of parameters canceling each other out as in the equation $\beta = -\alpha\gamma$. If we a priori rule out these unfaithful distributions (which have Lebesgue measure 0) then the conclusion must be that X does not cause Y.

Example 1. This example is the same as above but we drop the Normality assumption and state the setup in more generality. Let $V = (Z, X, Y)$ where Z represents all unobserved confounders, X is an exposure variable and Y is an outcome. We observe X and Y but not Z. Thus, $U = Z$ and $O = (X, Y)$. We further assume that the variables are in known time order with Z preceding X and X preceding Y. There are exactly eight possible DAG's for these variables as shown in Figure 2, all of which are subgraphs of G. $\mathcal{G}$ is this set of 8 DAGs. We have partitioned the DAGs into a set A with five DAG's and a set B of three DAG's. Assuming faithfulness, $X \amalg Y$ for the DAG's in the set B, and X and Y are dependent for the DAG's in A.

Suppose $X \amalg Y$ so that the true DAG G generating the data is in set B. Since there is no arrow from X to Y in any DAG in B, we then conclude that X does not cause Y. Suppose now that X is not independent of Y so the true DAG G is in set A. Since, in set A, some DAG's have an arrow from X to Y while others don't, we report:"no conclusion." Unlike most familiar statistical procedures, a key component of these causal procedures is that they can out output "no conclusion." This is formalized in Section 4. In the realistic setting where the population distribution of O is unknown, SGS suggest we perform an α-level test of the null hypothesis $X \amalg Y$ based on data O^n where the level may depend on n. If the test does not reject, we will accept $X \amalg Y$ and conclude X did not cause Y. We shall see that this test procedure can result in pointwise consistent but not uniformly consistent tests as defined below.

Example 2. Now suppose we have three time ordered variables (X, Y, Z). We want to know if Y causes Z. We allow for a confounding variable R between Y and Z and another possible confounding variable S between X and Z; see Figure 3. Hence, $V = (X, Y, Z, R, S)$. The decomposition into observables and unobservables is $O = (X, Y, Z)$ and $U = (R, S)$. We could also allow for a third confounding variable between X and Y but this is not crucial to the example.

Suppose we are given that

(i) X and Y are not independent, and not independent given Z

(ii) Y and Z are not independent and not independent given X

(iii) X and Z are not independent but are independent given Y

Assuming faithfulness, and using the d-separation rules of Pearl (1988), it follows that at least one arrow from R and one arrow from S must be removed. Moreover, the arrow from Y to Z must be present. We have thus inferred that Y causes Z. (The reader who is unfamiliar with the d-separation rules will not find this immediately obvious from looking at the DAG. d-separation is defined in the appendix.) In the settings when the population distribution of O is unknown, SGS suggests testing the joint null hypothesis (i)-(iii) with an α-level test, and if it does not reject, concluding that (i)-(iii) are true and thus that Y causes Z. Again, this test can be made to be pointwise but not uniformly consistent.

These examples give a sense of how causal inferences are made in this framework. See Spirtes, Glymour, and Scheines (1993) for the details of the general theory, which in general does not require either time order or the assumption of no latent variables. For clarity, the results in this paper focus on these last two paradigmatic cases. It is possible to extend the results beyond these cases.

4. CONSISTENT TESTS

To answer questions about the presence or absence of causal arrows from data in the SGS model, we need to use some test (or model search technique) for choosing between alternative causal models. In this section we review some basic facts about statistical tests.

Recall that G is a DAG for variables $V = (X_1, \ldots, X_k)$, $\mathcal{P}(G)$ are all probability distributions Markov to G and $\Omega(G) \subset \mathcal{P}(G)$ are the distributions that are faithful to G. Define

$$\mathcal{P}\mathcal{G} = \cup_{G \in \mathcal{G}} \{(P, G); P \in \mathcal{P}(G)\} \quad \text{and} \quad \Omega\mathcal{G} = \cup_{G \in \mathcal{G}} \{(P, G); P \in \Omega(G)\}.$$

Thus, (P, G) is in $\Omega\mathcal{G}$ if and only if P is faithful to some $G \in \mathcal{G}$. Note that $\Omega\mathcal{G}$ is not the Cartesian product of $\Omega(\mathcal{G})$ and $\mathcal{G}$. Recall also that $V = (O, U)$ where O is the subset of the variables of V that are observable and U are the remaining variables. For any $P \in \mathcal{P}(\mathcal{G})$, $\widetilde{P}$ denotes the marginal distribution O under P. We assume we have a random sample $O^n = (O_1, \ldots, O_n)$ from $\widetilde{P}$.

4.1. Consistent Tests. Let $\mathcal{O}^n = \mathcal{O} \times \cdots \times \mathcal{O}$ where $\mathcal{O}$ is the range of the random variable O. P^n denotes the n-fold product measure corresponding to P. Let $\theta = T(P, G)$ where T maps $\Omega\mathcal{G}$ into $\mathcal{R}$. We call θ a "parameter." In Section 4.2 we will formally define the "causal effect" which will be the parameter we are interested in. For now we let θ be an arbitrary parameter.

Let θ_0 be a fixed constant and consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. A test is a mapping $\phi : \mathcal{O}^n \to \{0, 1, 2\}$. The interpretation is $\phi(O^n) = 0$ means "choose H_0", $\phi(O^n) = 1$ means "choose H_1", $\phi(O^n) = 2$ means "make no decision." (We ignore randomized tests which make a random rather than a deterministic decision.) We shall be studying the asymptotic (large sample) properties of tests. Thus, suppose we specify a test ϕ_n for each sample size n and let $\phi = (\phi_1, \phi_2, \ldots)$ represent this sequence of tests. In what follows, all limits refer to the sample size n tending to ∞. When we refer to a test, we mean a sequence of tests ϕ. A test which always report "no decision" is not useful. To rule out such trivial cases, we make the following definition.

DEFINITION 1. A test ϕ is **non-trivial** if there exists some $P \in \Omega(\mathcal{G})$ and some $j \in \{0,1\}$ such that

$$\lim_n P^n(\phi_n(O^n) = j) = 1. \tag{1}$$

In other words, a test is non-trivial if there is at least one distribution P in the model such that, with probability tending to one, it eventually settles on a definite decision. Throughout the rest of this paper, we restrict attention to non-trivial tests. A test is consistent if, at least for large sample sizes, it does not report an incorrect decision. This is formalized as follows. Let

$$\begin{aligned}
\mathcal{P}_{\mathcal{G}0} &= \cup_{G\in\mathcal{G}}\{P \in \mathcal{P}(G); T(P,G) = \theta_0\} \\
\mathcal{P}_{\mathcal{G}1} &= \cup_{G\in\mathcal{G}}\{P \in \mathcal{P}(G); T(P,G) \neq \theta_0\} \\
\Omega_{\mathcal{G}0} &= \cup_{G\in\mathcal{G}}\{P \in \Omega(G); T(P,G) = \theta_0\} \\
\Omega_{\mathcal{G}1} &= \cup_{G\in\mathcal{G}}\{P \in \Omega(G); T(P,G) \neq \theta_0\}.
\end{aligned}$$

DEFINITION 2. A test ϕ is **pointwise consistent** if
(i) for every $P \in \Omega_{\mathcal{G}0}$, $P^n(\phi_n(O^n) = 1) \to 0$ and
(ii) for every $P \in \Omega_{\mathcal{G}1}$, $P^n(\phi_n(O^n) = 0) \to 0$.

Let $\mathcal{P}_{\mathcal{G}\delta} = \cup_{G\in\mathcal{G}}\{P \in \Omega(G); |T(P,G) - \theta_0| \geq \delta\}$, i.e. $\mathcal{P}_{\mathcal{G}\delta}$ contains distributions P compatible with a DAG G such that the treatment effect is at least δ away from θ_0.

DEFINITION 3. A test ϕ is **uniformly consistent** if
(i) $\sup_{P\in\Omega_{\mathcal{G}0}} P^n(\phi_n(O^n) = 1) \to 0$ and
(ii) for every $\delta > 0$, $\sup_{P\in\Omega_{\mathcal{G}\delta}} P^n(\phi_n(O^n) = 0) \to 0$.

The difference between a pointwise and uniformly consistent test is the presence of the suprema in the definitions. 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 and accepting the alternative 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 distribution that generates the data. That is, $n_0(\epsilon)$ is a function of the unknown P. Hence, there is no sample size n which guarantees that the probability of choosing the wrong hypothesis is less than ϵ, if P is unknown.

If a test is uniformly consistent then we can find $n_0(\epsilon)$ which does not depend on P such that $\sup_{P\in\Omega_{\mathcal{G}0}} P^n(\phi_n(O^n) = 1) \leq \epsilon$ 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 can 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 falsely accepting the null and rejecting the alternative when $|T(P, G)| > \delta$ is also bounded above by ϵ for all $n \geq n_0$, without knowledge of P or G.

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

4.2. Causal Effects and Hypotheses About Causal Graphs. In our earlier informal description of example 1, we constructed tests as follows. We began with a DAG G, and let $\mathcal{G}$ be the set of all subgraphs of G. We took A and B to be two sets of sub-graphs. Define $\Omega(A)$ to be all distributions faithful to some DAG in A and $\Omega(B)$ to be all distributions faithful to some DAG in B. We constructed A and B such that $\Omega(\mathcal{G}) = \Omega(A) \cup \Omega(B)$ and $\Omega(A) \cap \Omega(B) = \emptyset$. Finally, we tested A versus B by testing $P \in \Omega(A)$ versus $P \in \Omega(B)$. To put this in the terminology of the previous sub-section, deciding that $P \in \Omega(B)$ is like reporting $\phi_n(O^n) = 0$ while deciding $P \in \Omega(A)$ is like reporting $\phi_n(O^n) = 2$. At this point, it will clarify the discussion if we introduce the notion of a causal effect and then express our tests in terms of the causal effect parameter.

The causal effect of a variable X_i on another variable X_j can be thought of as the mean (or some other summary) of the distribution of X_j when X_i is set to a particular value x. To make this precise, let

$$p(v) = \prod_{r=1}^{k} p(x_r|pa_G(x_r))$$

be the joint density of V, where $PA_G(X_r)$ is the set of parents of X_r in G, and $pa_G(x_r)$ is some particular configuration of values of $PA_G(X_r)$. Define a new distribution $p_{G,X_i=x}(v)$ obtained by starting with the joint density $p(v)$ and replacing the factor $p(x_i|pa_G(x_i))$ with a distribution which is a point mass at x, i.e.

$$p_{G,X_i=x}(v) = \delta(x_i) \prod_{r\neq i} p(x_r|pa_G(x_r))$$

where $\delta(x_i) = 1$ if $x_i = x$ and $\delta(x_i) = 0$ otherwise.

Graphically, $p_{G,X_i=x}(x_j)$ has the following interpretation. Start with the graph G and create a new graph by breaking all arrows into X_i. Then, $p_{G,X_i=x}$ is the distribution in the new graph.

We regard $p_{G,X_i=x}$ as the density obtained by *setting* $X_i = x$. It should not be confused with $p(\cdot|X_i = x)$, which is the distribution when X_i is *observed* to be x. We call the distribution with density $p_{G,X_i=x}$ the **causal distribution**. Note that unlike $p(\cdot|X_i = x)$, $p_{G,X_i=x}$ depends upon both the joint distribution p, and on the graph G (i.e. two different data generating mechanisms represented by G_1 and G_2 respectively could generate the same distribution, but disagree on their respective causal distributions, $p_{G_1,X_i=x}$ and $p_{G_2,X_i=x}$).

Let

$$\tilde{p}_{G,X_i=x}(x_j) = \sum_{x_r;r\neq j} p_{G,X_i=x}(v)$$

be the marginal density for X_j obtained from $p_{G,X_i=x}(v)$.

The causal effect is usually defined to be some contrast functional of $\tilde{p}_{G,X_i=x}(x_j)$. In particular, assume that X_i is binary and define the causal effect θ as the mean of X_j when X_i is set to 1 minus the mean of X_j when X_i is set to 0. Specifically, if $\mathcal{X}_j$ denotes the possible values that X_j can take, then the **causal effect** is

$$\begin{aligned}
\theta &= \sum_{x_j\in\mathcal{X}_j} x_j \tilde{p}_{G,X_i=1}(x_j) - \sum_{x_j\in\mathcal{X}_j} x_j \tilde{p}_{G,X_i=0}(x_j) \\
&= E_{\tilde{P}_{G,X_i=1}}(X_j) - E_{\tilde{P}_{G,X_i=0}}(X_j).
\end{aligned}$$

The parameter θ is sometimes called the **treatment effect**. Since θ is a function of the joint distribution P and G we write $\theta = T(P, G)$. An alternative way to define the causal effect is by way of the theory of counterfactuals used by Rubin (1974) and Robins (1986, 1987, 1995, 1996). This approach leads to exactly the same formulas for the casual effects. A recent discussion the counterfactual approach is given in Dawid (2000).

When applied to example 1, the formula for the causal effect of X on Y turns out to be

$$\theta \equiv T(P) \equiv \int [E(Y|X = 1, Z = z) - E(Y|X = 0, Z = z)]dP(z). \tag{2}$$

In example 2, the formula for the causal effect of Y on Z is

$$\begin{aligned}
\theta &\equiv T(P) \\
&\equiv \int [E(Z|Y = 1, R = r, S = s) - E(Z|Y = 0, R = r, S = s)]dP(r, s).
\end{aligned}$$

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

In this section we discuss the non-existence of uniformly consistent tests for causal hypotheses in the two paradigmatic examples from Section 3. Recall that the first case involves a potential cause X, an outcome Y and a potential confounder Z. The second case involves a covariate X, a potential cause Y, an outcome Z and two potential confounders R and S. In each case we will take X and Y to be binary. The variables Z, R and S are discrete and take only finitely many values.

5.1. Case 1. The variables $V = (Z, X, Y)$ are assumed to be time ordered. The variable Z represents possible confounding variables. The question of interest is whether X causes Y i.e. is there an arrow from X to Y?

Recall that from observing X and Y to be nearly independent in a large sample, the conclusion from the informal reasoning in Section 3 is that X does not cause Y i.e. $\phi_n(O^n) = 0$. However, when X and Y were observed to be dependent, the outcome of the test was "no decision about causation" i.e. $\phi_n(O^n) = 2$. Let G be the complete DAG for V (with an arrow from X to Y) as in Figure 1. Let $\mathcal{G}$ be G and all sub-graphs of G. Following the discussion in Section 4, we test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

In this example, where the time order is given, the results about the existence of pointwise and uniformly consistent tests depend on whether or not Z is observed. In practice, we are mainly interested in the case where Z is unobserved. We include the case where Z is observed to make it clear exactly what is lost by the presence of the possibility of unobserved confounding. Note that when Z is observed we have $O = (Z, X, Y)$ and $U = \emptyset$ while if Z is unobserved then we have $O = (X, Y)$ and $U = Z$.

THEOREM 1. *Consider testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Given the Markov assumption and full Markov support, if Z is observed then there exist pointwise and uniformly consistent tests of this hypothesis. Given the Markov assumption, the faithfulness assumption, and full Markov support, if Z is not observed then there exist pointwise consistent tests but there are no uniformly consistent tests.*

Remark: The Theorem may be extended to testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ with $\theta_0 \neq 0$. In this case, when Z is unobserved, there is not even a non-trivial, pointwise consistent test.

The proof of Theorem 1 is in Section 10. (Related results may be found in Spirtes, Glymour, and Scheines, 1993.) Note that in the absence of the assumption that there are no latent variables, if faithfulness is not assumed then there are not even pointwise consistent tests of the hypothesis when $\theta_0 = 0$. This is because the independence of X and Y can always be explained either faithfully by no edge from X to Y and either no edge from Z to X or Y, (in which case the treatment effect is zero), or unfaithfully, (in which case the treatment effect is not zero).

Given the Markov and faithfulness assumptions, a time order, no latent variables, and that the population distribution P has full Markov support relative to the true causal graph G, P is compatible only with G. In that case, θ is a function of P alone. However, if one of the variables such as Z is not measured, then while the causal effect $\theta = T(P)$ is a function of P, only the marginal $\widetilde{P}$ can be consistently estimated from the data. In particular, we can construct P and Q so that their marginals are close in variation distance, but far apart in treatment effect, i.e. $\widetilde{P}$ and $\widetilde{Q}$ are nearly identical yet $|T(P) - T(Q)| > \delta > 0$. We can do this by taking Q to be faithful but arbitrarily close to being unfaithful.

It is worth recalling that there do exist uniformly consistent tests for testing associations. For example, a parameter that measures the association between X and Y is the risk difference $\alpha = Pr(Y = 1|X = 1) - Pr(Y = 1|X = 0)$. Another is the odds ratio

$$\beta = \frac{Pr(Y = 1|X = 1)Pr(Y = 0|X = 0)}{Pr(Y = 0|X = 1)Pr(Y = 1|X = 0)}.$$

PROPOSITION 1. *Let α and β be the parameters defined above. Let $O = (X, Y)$. For any α_0, there exists a uniformly consistent test for $H_0 : \alpha = \alpha_0$ versus $H_1 : \alpha \neq \alpha_0$. Similarly for β.*

This result follows since, for associations, we have that α and β are functions of P only through $\tilde{P}$.

5.2. Case 2. Now consider time ordered variables X, Y and Z and potential confounding variables R and S. We are interested in the causal effect of Y on Z. The complete directed acyclic graph to describe the model is given in Figure 3. Again, we take all the variables to be discrete. Recall that θ is now the causal effect of Y on Z as defined in Section 4.

THEOREM 2. *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Given the Markov assumption and full Markov support, if R and S are observed, there exist pointwise and uniformly consistent tests for this hypothesis. Given the Markov assumption, the faithfulness assumption, and full Markov support, if R and S are unobserved, there exist pointwise consistent tests for this hypothesis but there do not exist uniformly consistent tests.*

Since the proof of Theorem 2 is similar to the proof of Theorem 1, it is omitted.

Remark: Note that, in contrast to the situation in Theorem 1, there do exist non-trivial, pointwise consistent tests for the case where $\theta_0 \neq 0$. This may seem surprising. However, in this case, the null hypothesis is equivalent to the intersection of three associational nulls – H_{01} : X and Y are *not* independent, H_{02} : Y and Z are *not* independent, H_{03} : X and Z are independent given Y – and there are pointwise consistent tests for each of these hypotheses.

Remark. In a randomized experiment, it is possible to construct uniformly consistent tests of causal hypotheses. To see this, return to case 1 of Section 5.1. Randomization breaks the arrow from Z to X in Figure 1. It then can be shown that $\theta = \alpha = Pr(Y = 1|X = 1) - Pr(Y = 1|X = 0)$ and as noted above, there exists uniformly consistent tests for this risk difference α.

6. 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 of causal effects. We shall see that causal effects are not uniformly estimable in the presence of unobserved confounders for the same reasons that uniformly consistent tests do not exist.

6.1. Estimability of Causal Effects. An **estimator** is a sequence of functions $\hat{\theta}_n$ of $O^n = (O_1, \ldots, O_n)$.

An estimator is consistent if, for every $(P, G) \in \Omega\mathcal{G}$, and every $\epsilon > 0$, $P^n(|T(P, G) - \hat{\theta}_n| > \epsilon) \to 0$. An estimator $\hat{\theta}_n$ is **uniformly consistent** if, for every $\epsilon > 0$,

$$\sup_{(P,G)\in\Omega\mathcal{G}} P^n(|\hat{\theta}_n - T(P, G)| > \epsilon) \to 0.$$

To discuss estimation in the SGS framework, we need to allow for a more general notion of an estimator. Recall that their tests sometimes return "no conclusion" which is what allows the tests to be pointwise consistent. Similarly, for estimation, we should allow the possibility that the answer is sometimes "no estimate is provided." Formally, we define a **generalized estimator** $\hat{\theta}_n$ to be any measurable mapping from $\mathcal{O}^n$ into the Borel sets of the real line. In particular, when $\hat{\theta}_n(O^n) = \mathcal{R}$ (as opposed to a point) this corresponds to "no estimate is provided." To define consistency for generalized estimators we need to define the distance between the estimator and the true value. Given a set A and a point b define $\tau(A, b) = \inf_{a\in A} d(a, b)$. If $A = \{a\}$ consists of a single point, then $\tau(A, b) = |a - b|$, the usual Euclidean distance. A generalized estimator is **consistent** at (P, G) if for every $\epsilon > 0$, $P^n(\tau(\hat{\theta}_n, T(P, G)) > \epsilon) \to 0$. A generalized estimator $\hat{\theta}_n$ is **uniformly consistent** if, for every $\epsilon > 0$,

$$\sup_{(P,G)\in\Omega\mathcal{G}} P^n(\tau(\hat{\theta}_n, T(P, G)) > \epsilon) \to 0.$$

A generalized estimator $\hat{\theta}_n$ is **non-trivial** if there exists $P \in \Omega(\mathcal{G})$ such that

$$P^\infty(A_n^c \text{ finitely often }) = 1$$

where A_n is the event that $\hat{\theta}_n$ consists of a single point. In other words, $\hat{\theta}_n$ is non-trivial if there is at least one P for which $\hat{\theta}_n$ eventually becomes a point.

THEOREM 3. *Given the Markov assumption and full Markov support, in case 1, if Z is observed, and in case 2, if R and S are observed, and there are no unobserved variables, there is a uniformly consistent estimator of θ. Given the Markov assumption, the faithfulness assumption, and full Markov support, in case 1, if Z is not observed, and in case 2, if R and S not are observed, there is no uniformly consistent estimator of θ.*

The proof is very similar to the proof of Theorems 1 and 2 and is omitted. Some insight can be gained into the lack of consistency in the following way. Let H be the map that takes P into $\tilde{P}$. The parameter θ is **identified at** P if $\theta(P)$ is the same for all $H^{-1}(\tilde{P})$. It is well known that if a parameter is not identified for any P then it cannot be consistently estimated. In our case, the situation is a bit subtle. If the parameter were defined on $\mathcal{P} \equiv \mathcal{P}(\mathcal{G})$ instead of $\Omega \equiv \Omega(\mathcal{G})$, then it would follow immediately that θ is not identifiable for any P and hence not consistently estimable. One might hope that by restricting ourselves to Ω instead of $\mathcal{P}$ might help since, for example, in Example 1, if $\tilde{P}$ is a distribution with X independent of Y then θ is identified at P and is equal to 0. However, it can be shown that Ω is dense in $\mathcal{P}$, so uniform consistent estimation is still precluded. As with testing, there are uniformly consistent estimators of α and β.

6.2. Confidence Intervals. Consider a collection of maps

$$\mathcal{C} = \{I_{\alpha,n};\ n = 1, 2, \ldots, \alpha \in [0,1]\}$$

where each $I_{\alpha,n}$ takes $\mathcal{O}^n$ into the Borel sets of $\mathcal{R}$. If

$$\text{liminf}_n \inf_{(P,G)\in\Omega\mathcal{G}} P^n(T(P,G) \in I_{\alpha,n}(O^n)) \geq 1-\alpha$$

then we call $\mathcal{C}$ a **confidence map** and we call $I_{\alpha,n}$ a $1-\alpha$ asymptotic, confidence region. A confidence map $\mathcal{C}$ is **consistent** if it eventually omits all false values i.e. if for every $\delta > 0$, for every $\alpha \in [0,1]$,

$$\sup_{\{(P,G),(Q,H)\}\in\Omega\mathcal{G};|(T(Q,H)-T(P,G))|>\delta} P^n(T(Q,H) \in I_{\alpha,n}(O^n)) \to 0.$$

THEOREM 4. *Given the Markov assumption, the faithfulness assumption, and full Markov support, in example 1, if Z is unobserved then there does not exist a consistent, confidence map for θ. Similarly, in example 2, if R and S are unobserved.*

7. BAYESIAN INFERENCE

So far, we have concentrated on frequentist inference. Robins and Wasserman (1999) discuss related Bayesian results. This section gives a brief summary of the Bayesian results and how they relate to the frequentist results; see Robins and Wasserman (1999) for a full account, and Spirtes et al. (1999) for a discussion. The goal of the section is to elucidate the behavior of the causal test as a function of the prior distribution. We focus on Example 1.

There are weaker and stronger senses of Bayesian convergence, which are analogous to pointwise and uniform consistency. In the weaker sense (which uses a fixed prior) of Bayesian convergence, the procedures described in Spirtes, Glymour, and Scheines (1993) asymptotically converge to the truth for any prior over the parameters which is absolutely continuous with Lebesgue measure and the true causal DAG has non-zero probability; in the stronger sense of Bayesian convergence (which uses a prior that changes with sample size), whether or not these procedures asymptotically converge to the truth is more sensitive to the prior (and in particular depends upon how probable confounding is relative to the sample size.)

There are 8 sub-graphs in Example 1 (see Figure 2). Denote these sub-graphs $G_1, \ldots, G_8$. Let γ_j represent the parameters of the joint distribution for G_j. For example, in the complete graph G_4, we could define the parameters γ_4 by

$$\gamma_4 = \{\gamma_{4z}, \gamma_{4xz}, \gamma_{4yxz}; z \in \mathcal{Z}, x, y \in \{0,1\}\}$$

where $\gamma_{4z} = Pr(Z = z)$, $\gamma_{4xz} = Pr(X = x|Z = z)$ and $\gamma_{4yxz} = Pr(Y = y|X = x, Z = z)$.

For a Bayesian analysis we put a prior on the sub-graphs and on the parameters of each sub-graph. We can write the prior as $\pi(G_j)\pi(\gamma_j|G_j)$. We assume that the prior on γ_j is smooth (i.e. absolutely continuous with respect to Lebesgue measure with bounded density). After observing the data O^n, Bayes' theorem gives us a posterior probability $\pi(G_j|O^n)$ for each possible sub-graph. We can then test for the presence of a causal effect by finding the posterior odds B_n of "X causes Y" versus "X does not cause Y":

$$B_n = \frac{Pr(M|O^n)}{Pr(M^c|O^n)} = \frac{Pr(H_1|O^n)}{Pr(H_0|O^n)}$$

where M is the set of all graph's G in which there is an arrow from X to Y i.e. M corresponds to H_1. The number B_n tells us the strength of evidence in favor of the hypothesis of causation. In testing consistently estimable parameters such as measures of association, it is typically the case that B_n will tend to 0 or infinity in probability as the sample size increases and thus B_n will be decisive.

Let $\mathcal{H}$ be all subgraphs where Z does not have arrows into both X and Y. Let us compare the Bayesian and frequentist analysis. In contrast to the Bayesian analysis, the frequentist analysis is a worst case analysis: for each sample size n we can always choose a P_n which is (i) faithful but arbitrarily close to unfaithful, (ii) which has a large causal effect, and (iii) whose marginal makes X and Y nearly independent. The existence of such a single such P_n precludes a uniformly consistent test. However, as n grows, the offending P_n must be closer and closer to being unfaithful. In other words, the set of offending P_n gets smaller as n increases. Let $\mathcal{P}_n$ be the set of offending P_n's. With fixed prior mass on each subgraph and a smooth, fixed prior density on the parameters of each sub-graph, (the weaker sense of Bayesian convergence) we see that $\pi(\mathcal{P}_n) \to 0$. Thus, these distributions play a lesser and lesser role as n increases which is why $B_n \to 0$ in probability (in the cases mentioned above).

There is a stronger sense of Bayesian convergence, which more closely links the asymptotic results with finite sample sizes. In this stronger sense of Bayesian convergence, instead of using a fixed prior, the prior mass on the graphs changes with sample size n. This is only meant to reflect the fact that the prior on $\mathcal{H}$ can be small relative to sample size. We do not literally envisage the prior changing with n. In this stronger sense of Bayesian convergence, some prior distributions will lead to tests that are pointwise consistent while others will lead to trivial tests. $\mathcal{H}$ can be thought of as the hypothesis that there is no unobserved confounding. Robins and Wasserman (1999) showed that, under weak conditions, if $\pi(\mathcal{H}) = o(n^{-1/2})$, then B_n will stay bounded (in probability) away from 0 and infinity and hence not be decisive. Thus, if one's prior belief that there is no confounding (relative to the sample size) is small, the Bayes test provides no decision. On the other hand, if there were reason to believe that there might not be unobserved confounding, then we would set $\pi(\mathcal{H})$ to be not small relative to sample size (for example, put fixed positive mass on each sub-graph). In this case, when $X \amalg Y$, $B_n \to 0$ in probability.

As pointed out previously, we can always choose a P_n which is (i) faithful but arbitrarily close to unfaithful, (ii) which has a large causal effect and (iii) whose marginal makes X and Y nearly independent. With a fixed prior, as sample size grows these distributions play a lesser and lesser role But, if, as sample size increases, we pile more probability near the unfaithful distributions or we put less and less mass on the sub-graphs with no confounders, and if we capture this by using a prior π_n that depends on n, then $\pi_n(\mathcal{P}_n)$ does not tend to 0 and the Bayes test makes no decision. Which priors are more reasonable depends on subject matter knowledge. See Robins and Wasserman (1999), Spirtes et al. (1999), and Spirtes, Glymour, and Scheines (2000) for a discussion of the role of the prior. There are parallel results for point and interval estimation but we shall not pursue the details here.

8. INFERRING TIME ORDER

Throughout this paper we have assumed that the variables have known time order. The lack of uniform consistency arose because of the presence of potential unobserved confounding variables. The SGS procedures that we have discussed also permit one to infer the time order of the variables in some cases. The purpose of this section is to show that, for the problem of inferring time order, there again do not exist uniformly consistent procedures. But in this case, the lack of uniformity does not even require unobserved confounding variables.

Suppose we have four, binary, random variables $V = (X_1, \ldots, X_4)$. Consider the two graphs in Figure 4. Suppose we are interested in the causal effect of X_3 on X_4. In the first graph, it can be shown that the causal effect is $\theta = Pr(X_4 = 1|X_3 = 1) - Pr(X_4 = 1|X_3 = 0)$. In the second graph, θ is identically 0. This follows because the arrow points from X_4 to X_3 so manipulating X_3 will have no effect on X_4. Notice that getting the direction of the arrow wrong thus implies getting the value of θ wrong. So we will cast questions about getting the directions of arrow correctly as questions about the causal effect θ.

Is it possible to distinguish these two DAG's? Suppose the following independence relationships were known: $X_1 \amalg X_2$, $\{X_1, X_2\} \amalg X_4 | X_3$ and that no other independence relationships not entailed by these hold. Assume there are no unobserved confounding variables. Then only the first DAG is consistent with these relationships assuming faithfulness, and there are pointwise consistent tests of the effect of X_3 on X_4. However, there are no uniformly consistent tests of whether the effect of X_3 on $X_4 = \delta \neq 0$.

For inferring time order, G is also unknown. Define $\mathcal{G}$ to be all DAG's for $X_1, \ldots, X_4$.

THEOREM 5. *Suppose we have n i.i.d. observations $V^n = (V_1, \ldots, V_n)$ and that there are no unobserved confounding variables. Given the Markov assumption, the faithfulness assumption, and full Markov support, in the above example, there are pointwise consistent tests but no uniformly consistent, test of $H_0 : \theta = \delta$ versus $H_1 : \theta \neq \delta$. Similarly, there is no uniformly consistent point estimator and there is no consistent $1 - \alpha$ confidence interval for θ.*

Remark: Note that the case, $H_0 : \theta = 0$ is not included in Theorem 5. When trying to find a causal time order with no confounders, the value 0 plays a special role. To see this, consider the simple case of two variables X and Y. Suppose we know that there are no confounding variables but we do not know the time order so we are trying to decide if $X \rightarrow Y$ or $Y \rightarrow X$. In particular, let θ be the causal effect of X on Y. Clearly, it is not possible to consistently test whether $\theta = \delta$, for $\delta \neq 0$. Any correlation between X and Y can be explained equally well by the graph $X \to Y$ or the graph $Y \to X$. However, we can test whether $\theta = 0$ merely by testing for an association between X and Y. Even without time order, the only explanation for no association between X and Y is that $\theta = 0$.

9. REMARKS

Many results in statistical inference are asymptotic, that is, they hold true for large sample sizes. In using asymptotic results, it is important that the large sample results can be approximated with finite samples. One way to ensure this is true is to insist on our tests and estimates being uniformly consistent. We have shown that causal procedures based on the Markov and Faithfulness assumptions are not uniformly consistent. The *same* problem affects model search or model averaging methods. In these methods, each sub-graph obtained by deleting some set of arrows is regarded as a separate model. Then a model search technique (such as Bayesian model selection, BIC, AIC etc.) is used to find a best graph or to average over graphs. These methods will also fail to have uniform consistency properties, as will any variable selection technique that attempts to eliminate variables with small treatment effects on the basis of the observed data.

This leaves open the question of what to do when analyzing observational studies. Each of the suggested strategies is controversial. The problems stem from unobserved confounding or unknown time order. Given a time order, one suggestion is to follow standard epidemiological advice and measure as many confounders as possible. But even then we cannot be sure that there are not further unobserved confounding variables. Further, conditioning on inappropriate variables can introduce bias into the procedures. (Spirtes et al 1998, Greenland, Pearl and Robins, 1999). A second possible approach is performing a sensitivity analysis in which we quantify how the estimates and tests change as a function of the amount of unobserved confounding. Examples of sensitivity analysis can be found in Rosenbaum (1993), Robins et. al (1999), Manski (1990, 1995), and Balke and Pearl (1997). A third strategy is to continue to pursue search strategies which are not uniformly consistent, but do satisfy weaker consistency conditions (e.g. Bayesian consistency conditions.)

10. PROOFS

In the proofs, we make use of the total variation distance. If P and Q are two probability measures, the the total variation distance between P and Q is defined to be $d(P, Q) = \sup_A |P(A) -$

$Q(A)|$ the supremum being over all measurable sets. If P and Q live on a finite sample space $S = \{s_1, \ldots, s_m\}$ then it can be shown that $d(P,Q) = (1/2)\sum_{j=1}^{m} |p_j - q_j|$ where $p_j = P(\{s_j\})$ and $q_j = Q(\{s_j\})$. Furthermore, $d(P,Q)$ is a continuous function of the $p_j's$ and the $q_j's$. For any vector $a = (a_1, \ldots, a_K)$ we define $||a||_\infty = \max_j |a_j|$.

Proof of Theorem 1

We only prove the case $\theta_0 = 0$ which is the more difficult case. First suppose that Z is observed. We construct a test that is both uniformly and pointwise consistent. Let $\Delta = \Delta(P) = (\beta_{000}, \ldots, \beta_{11B})$, where $\beta_{rst} = Pr(X = r, Y = s, Z = t)$ for $r, s \in \{0,1\}$ and $t \in \{0, \ldots, B\}$. In this example, because the time order is given, and there are no latent variables, the treatment effect is completely determined by Δ. Let $\hat{\Delta} = (\hat{\beta}_{000}, \ldots, \hat{\beta}_{11B})$ where $\hat{\beta}_{rst}$ is the observed sample proportion corresponding to β_{rst}, i.e. $\hat{\beta}_{rst} = \#\{(X_i, Y_i, Z_i); X_i = r, Y_i = s, Z_i = t\}/n$. Let $\epsilon_n = \sqrt{\log n/n}$. It follows from Hoeffding's inequality that

$$\sup_{P \in \Omega(\mathcal{G})} P^n(||\hat{\Delta} - \Delta(P)||_\infty > \epsilon_n) \to 0.$$

Define $\hat{\theta} = T(\hat{\Delta})$. Since θ is a continuous function of Δ,

$$\sup_{P \in \Omega(\mathcal{G})} P^n(|\hat{\theta} - \theta(P)| > \epsilon_n) \to 0. \tag{3}$$

Define $\phi_n(O^n) = 1$ if $|\hat{\theta}| > \epsilon_n$ and $\phi_n(O^n) = 0$ otherwise. It follows from (3) that this test is uniformly and pointwise consistent.

Now assume Z is not observed. First we construct a pointwise consistent test. Define $\alpha = Pr(Y = 1|X = 1) - Pr(Y = 1|X = 0)$ and let

$$\hat{\alpha} = n^{-1}\left[\frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i} - \frac{\sum_{i=1}^n Y_i(1 - X_i)}{\sum_{i=1}^n (1 - X_i)}\right].$$

Again, we have that $\sup_{P \in \Omega(\mathcal{G})} P^n(|\hat{\alpha} - \alpha(P)| > \epsilon_n) \to 0$. Define $\phi_n(O^n) = 0$ if $|\hat{\alpha}| < \epsilon_n$ and $\phi_n(O^n) = 2$ otherwise.

If $\alpha(P) = 0$ then $P^n(\phi_n(O^n) = 0) \to 1$ so the test is non-trivial. Now we show that the test is pointwise consistent. Suppose $P \in \Omega_{\mathcal{G}0}$. Then P is faithful to one of the DAG's in B (figure 2) or the first DAG in A (DAG (i) in Figure 2.) Thus $\Omega_{\mathcal{G}0} = \Omega_B \cup \Omega^*$ where Ω_B is the set of distributions faithful to some DAG in A and Ω^* is the set of distributions faithful to the first DAG in A (DAG (i) in Figure 2.) Suppose that $P \in \Omega_B$. Since $\alpha(P) = 0$ for every $P \in \Omega_B$, and from the uniform convergence of $\hat{\alpha}$, we deduce that $P^n(\phi_n(O^n) = 1) \to 0$. If $P \in \Omega^*$, then $\alpha(P) \neq 0$ and hence $P^n(\phi_n(O^n) = 2) \to 1$. Thus, $P^n(\phi_n(O^n) = 1) \to 0$ for all $P \in \Omega_{\mathcal{G}0}$. If $P \in \Omega_{\mathcal{G}1}$ and P is faithful, then $\alpha(P) \neq 0$. Hence, from the uniform consistency of $\hat{\alpha}$ we see that $P^n(\phi_n(O^n) = 2) \to 1$. This establishes the pointwise consistency of the test.

Now we show that no test can be uniformly consistent. Suppose there is a non-trivial, uniformly consistent test. We will show that this leads to a contradiction. Since ϕ is non-trivial, (1) holds for some P. Suppose $T(P) = \delta$ for some $\delta \neq 0$, so $P \in \Omega_{\mathcal{G}\delta}$ and $P^n(\phi_n(O^n) = 1) \to 1$ as $n \to \infty$. Let $p(x, y, z)$ denote the density of P. Let G be DAG (i) in Figure 2. Then $\mathcal{P}(G)$ is all Q whose densities satisfy $q(x, y, z) = q(z)q(x|z)q(y|z)$. Note that $Q \in \mathcal{P}(G)$ implies that $T(Q) = 0$ if Q has full Markov support. If $\tilde{P}$ has full support, let $P^*(X, Y) = \tilde{P}$, and otherwise let $P^*(X, Y)$ be a distribution with full support such that $d(\tilde{P}, P^*) < \epsilon$.

There exists a $Q \in \mathcal{P}(G)$ such that Q has full Markov support, is faithful to G, $\tilde{Q} = P^*$, and $T(Q, G) = 0$. To see this, define $\alpha_{rs} = P^*(X = r, Y = s)$ for $r, s \in \{0, 1\}$. Define a distribution Q as follows. Let $Q(Z = 0) = \alpha_{00}$, $Q(Z = 1) = \alpha_{01}$, $Q(Z = 2) = \alpha_{10}$, and $Q(Z = 3) = \alpha_{11}$ Also, let

$$\begin{aligned}
Q(X = 0, Y = 0|Z = 0) &= 1 \\
Q(X = 0, Y = 1|Z = 1) &= 1 \\
Q(X = 1, Y = 0|Z = 2) &= 1 \\
Q(X = 1, Y = 1|Z = 3) &= 1.
\end{aligned}$$

Let q be the corresponding density. It can be verified directly that $q(x, y, z) = q(z)q(x|z)q(y|z)$ so that $Q \in \mathcal{Q}$, has full Markov support, is faithful to G, $\tilde{Q} = P^*$, and $T(Q, G) = 0$. Hence $Q \in \Omega_{\mathcal{G}0}$. Since the test depends only on O^n and since $\tilde{Q} = P^*$, we have that

$$\begin{aligned}
\sup_{R \in \Omega_{\mathcal{G}0}} R^n(\phi_n(O^n) = 1) &\geq Q^n(\phi_n(O^n) = 1) \\
&= \tilde{Q}^n(\phi_n(O^n) = 1) \\
&= P^{*n}(\phi_n(O^n) = 1) \\
&> \tilde{P}^n(\phi_n(O^n) = 1) - \epsilon \\
&= P^n(\phi_n(O^n) = 1) - \epsilon \\
&\to 1 - \epsilon.
\end{aligned}$$

Hence, $\sup_{R \in \Omega_{\mathcal{G}0}} R^n(\phi_n(O^n) = 0)$ does not tend to 0 as required for a uniformly consistent test.

Now suppose instead that $P \in \Omega_{\mathcal{G}0}$. Thus, $P^n(\phi_n(O^n) = 0) \to 1$. If $X \amalg Y$ in $\tilde{P}$, let $\delta = P(Y = 0)P(Y = 1)/4$, and otherwise let $\delta = [P(Y = 1|X = 1) - P(Y = 1|X = 0)]/4$. For every ϵ, there exists a $Q \in \mathcal{P}_{\mathcal{G}\delta}$ such that $d(\tilde{P}^n, \tilde{Q}^n) < \epsilon/2$. This fact is proved in Proposition 2 below. However, Q might not be faithful. Since T and $d(\cdot,\cdot)$ are both continuous functions of P, there exists a faithful $Q_n \in \Omega_{\mathcal{G}\delta}$ with full Markov support such that (i) $|T(Q_n)| > \delta$, (ii) $d(Q_n^n, Q^n) < \epsilon/2$. It follows that $d(\tilde{Q}_n^n, \tilde{Q}^n) < \epsilon/2$. Hence,

$$\begin{aligned}
\sup_{R \in \Omega_{\mathcal{G}\delta/2}} R^n(\phi_n(O^n) = 0) &\geq Q_n^n(\phi_n(O^n) = 0) \\
&> Q^n(\phi_n(O^n) = 0) - \epsilon/2 \\
&= \tilde{Q}^n(\phi_n(O^n) = 0) - \epsilon/2 \\
&> \tilde{P}^n(\phi_n(O^n) = 0) - \epsilon/2 - \epsilon/2 \\
&= P^n(\phi_n(O^n) = 0) - \epsilon \\
&\to 1 - \epsilon.
\end{aligned}$$

Since this is true for any $\epsilon > 0$ we conclude that

$$\sup_{R \in \Omega_{\mathcal{G}\delta}} R^n(\phi_n(O^n) = 0) \to 1$$

and hence, the test is not uniformly consistent. $\diamond$

PROPOSITION 2. In Example 1, let $O = (X, Y)$. Consider $P \in \Omega_{\mathcal{G}0}$. If $X \amalg Y$ under P, let $\delta = (P(Y = 1)P(Y = 0))/4$; otherwise let $\delta = [P(Y = 1|X = 1) - P(Y = 1|X = 0)]/4$. For every $\epsilon > 0$ there exists a $Q \in \mathcal{P}_{\mathcal{G}\delta}$ such that $d(\tilde{P}, \tilde{Q}) < \epsilon/2$.

PROOF. First suppose that $p(x, y, z) = p(z)p(x|z)p(y|z)$. Thus, P is faithful to DAG (i) in B in Figure 2. It follows that $\alpha(P) \neq 0$ (i.e. X and Y are not marginally independent). If $\tilde{P}$ has full support, let $P^*(X, Y) = \tilde{P}$, and otherwise let $P^*(X, Y)$ be a distribution with full support such that $d(\tilde{P}, P^*) < \epsilon/2$, and $P^*(Y = 1|X = 1) - P^*(Y = 1|X = 0) > (P(Y = 1|X = 1) - P(Y = 1|X = 0))/2$.

Define Q as follows. Let $Q(Z = 0) = 1/2$ and let $Q(X = x, Y = y|Z = z) = \tilde{P}^*(X = x, Y = y)$ for $z = 0, 1$. By construction, $\tilde{Q} = \tilde{P}^*$. Let G be DAG (iv) in Figure 2. Using equation 2, $T(Q, G) = P^*(Y = 1|X = 1) - P^*(Y = 1|X = 0) > (P(Y = 1|X = 1) - P(Y = 1|X = 0))/2 > \delta$. Because $d(\tilde{P}, P^*) < \epsilon/2$ and $\tilde{Q} = P^*$, $d(\tilde{P}, \tilde{Q}) < \epsilon/2$.

For all other $P \in \Omega_{\mathcal{G}0}$ we have that $\alpha(P) = 0$ (X and Y are marginally independent) so the above strategy does not work. Instead we proceed as follows. Let $P^*(X, Y)$ be a distribution with full support such that $d(\tilde{P}, P^*) < \epsilon/2$, $X \amalg Y$ under P^*, and $P^*(Y = 0)P^*(Y = 1) > P(Y = 0)P(Y = 1)/2$.

Let $a = P^*(X = 0)$ and $b = P^*(Y = 0)$, and c and d be two reals between 0 and 1. Since X and Y are marginally independent under P^*, a and b completely determine P^*. Define Q as follows:

$$\begin{aligned}
Q(X = 0) &= a \\
Q(Y = 0) &= b \\
Q(Z = 0|X = 0, Y = 0) &= 1 - c \\
Q(Z = 0|X = 0, Y = 1) &= 0 \\
Q(Z = 0|X = 1, Y = 0) &= 1 - d \\
Q(Z = 0|X = 1, Y = 1) &= 0
\end{aligned}$$

and $q(x, y, z) = q(x)q(y)q(z|x, y)$ where q is the density of Q.

By inspection, $\tilde{Q} = \tilde{P^*}$. Q has the following properties:

$$\begin{aligned}
Q(Z = 1) &\geq (1 - b) \\
Q(Y = 1|X = 0, Z = 0) &= 0 \\
Q(Y = 1|X = 1, Z = 0) &= 0 \\
Q(Y = 1|X = 0, Z = 1) &= \frac{1 - b}{1 - b + cb} \\
Q(Y = 1|X = 1, Z = 1) &= \frac{1 - b}{1 - b + db}.
\end{aligned}$$

As $d \to 0$, $Q(Y = 1|X = 1, Z = 1) \to 1$, and as $c \to 1$, $Q(Y = 1|X = 0, Z = 1) \to (1 - b)$. Let G be DAG (iv) in Figure 2. Then, by choosing the proper values for c and d, and using equation 2, $T(Q, G) > b(1 - b)/2 = P^*(Y = 0)P^*(Y = 1)/2 > P(Y = 0)P(Y = 1)/4 = \delta$. Because $d(\tilde{P}, P^*) < \epsilon/2$ and $\tilde{Q} = P^*$, $d(\tilde{P}, \tilde{Q}) < \epsilon/2$. $\diamond$

Proof of Theorem 4.

Suppose a consistent confidence interval exists. Define a test as follows. Let α_n be a decreasing sequence such that $\alpha_n \to 0$ as $n \to \infty$. Let $\phi_n(O^n) = 1$ if $0 \notin I_{\alpha_n,n}$ and $\phi_n(O^n) = 0$ if $0 \in I_{\alpha_n,n}$. It is easy to see that this defines a uniformly consistent test which contradicts Theorem 1. $\diamond$

Proof of Theorem 5.

Let P have density p such that $p(v) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)$. It follows that $\theta = T(P, G_1) = Pr(X_4 = 1|X_3 = 1) - Pr(X_4 = 1|X_3 = 0)$. Moreover, we can choose such a p so that

$|\theta| \geq \delta$ for some $\delta > 0$. Fix $\epsilon > 0$. It is possible to choose Q_n with density q_n such that Q_n is Markov and Faithful to the second graph G_2 in Figure 4 and such that $d(P^n, Q_n^n) < \epsilon$. Note that $T(Q_n, G_2) = 0$. The remainder of the proof is similar to the previous proofs. $\diamond$

APPENDIX

Definition of D-Separation

Let a **directed path** D in a DAG G be a sequence of distinct vertices $V_1 \cdots V_n$ such that for $1 \leq i < n$, there is an edge $V_i \to V_{i+1}$ in G. A vertex X is an **ancestor** of Y if $X = Y$ or there is a directed path from X to Y. Let an **undirected path** U in a DAG be a sequence of distinct vertices $V_1 \cdots V_n$ such that for $1 \leq i < n$, there is either an edge $V_i \to V_{i+1}$ or $V_i \leftarrow V_{i+1}$ in G. V_i is a **collider** on U if there are edges $V_{i-1} \to V_i \leftarrow V_{i+1}$ on U; otherwise V_i is a **non − collider** on U. For three disjoint sets of vertices X, Y, and Z, X is **d − separated** from Y given Z if there is no undirected path U from a vertex in X to a vertex in Y such that every collider on U is an ancestor of a member of Z, and no non-collider on U is in Z. Pearl (1988) states that if $P \in \mathcal{P}(G)$ and X is d-separated from Y given Z then $X \amalg Y | Z$.

Glossary of Notation

Let $\mathcal{G}$ be a set of DAG's and let G denote a DAG in $\mathcal{G}$.

$$\begin{aligned}
\mathcal{P}(G) &= \text{All distributions Markov to } G \\
\Omega(G) &= \text{All distributions Faithful to } G \\
\mathcal{P}(\mathcal{G}) &= \cup_{G \in \mathcal{G}} \mathcal{P}(G) \\
\Omega(\mathcal{G}) &= \cup_{G \in \mathcal{G}} \Omega(G) \\
\mathcal{P}_{\mathcal{G}0} &= \cup_{G \in \mathcal{G}} \{P \in \mathcal{P}(G); T(P,G) = \theta_0\} \\
\mathcal{P}_{\mathcal{G}1} &= \cup_{G \in \mathcal{G}} \{P \in \mathcal{P}(G); T(P,G) \neq \theta_0\} \\
\Omega_{\mathcal{G}0} &= \cup_{G \in \mathcal{G}} \{P \in \Omega(G); T(P,G) = \theta_0\} \\
\Omega_{\mathcal{G}1} &= \cup_{G \in \mathcal{G}} \{P \in \Omega(G); T(P,G) \neq \theta_0\} \\
\mathcal{P}_{\mathcal{G}\delta} &= \cup_{G \in \mathcal{G}} \{P \in \mathcal{P}(G); |T(P,G) - \theta_0| > \delta\} \\
\Omega_{\mathcal{G}\delta} &= \cup_{G \in \mathcal{G}} \{P \in \Omega(G); |T(P,G) - \theta_0| > \delta\} \\
\mathcal{P}\mathcal{G} &= \cup_{G \in \mathcal{G}} \{(P,G); P \in \mathcal{P}(G)\} \\
\Omega\mathcal{G} &= \cup_{G \in \mathcal{G}} \{(P,G); P \in \Omega(G)\}.
\end{aligned}$$

REFERENCES

Balke, A. and Pearl, J. (1997). Bounds on Treatment Effects From Studies With Imperfect Compliance. *Journal of the American Statistical Society*, 92, 1171-1176.

Bollen, K. (1989). *Structural Equations with Latent Variables.* New York: Wiley.

Dawid, A. P. (2000). Causal inference without counterfactuals (with discussion), *Journal of the American Statistical Society*, 95, 407-448.

Donoho, D. (1988). One-sided inference about functionals of a density. *The Annals of Statistics*, **16**, 1390-1420.

Greenland, S., Pearl, J. and Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10, 37-48.

Lauritzen, S. (1996). *Graphical Models.* Clarendon Press: Oxford.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, **1**, 38-53.

Manski, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. (1995). *Identification Problems in the Social Sciences,* Harvard University Press: Cambridge.

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

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.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the*

12th Conference on Uncertainty in Artificial Intelligence, Portland, Oregon, 1996. E.Horvitz and F.Jensen (eds.), Morgan Kaufmann, San Francisco, CA.

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. (1995). “Discussion of ‘Causal Diagrams for empirical research’ by J. Pearl,” *Biometrika*, 82, 695-698.

Robins, J.M. (1997). Causal Inference from Complex Longitudinal Data. In *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120),* M. Berkane, Editor. NY: Springer Verlag, pp. 69-117.

Robins, J.M. and Wasserman, L. (1999). On the impossibility of inferring causation from association without background knowledge. In: *Computation, Causation, and Discovery.* Eds. C. Glymour and G. Cooper. Cambridge, MA: The MIT Press.

Rosenbaum, P. (1993). *Observational Studies.* Springer: New York.

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

Spirtes, P. (1992). Building causal graphs from statistical data in the presence of latent variables. *Proceedings of the IX International Congress on Logic and Methodology.* Uppsala, Sweden.

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

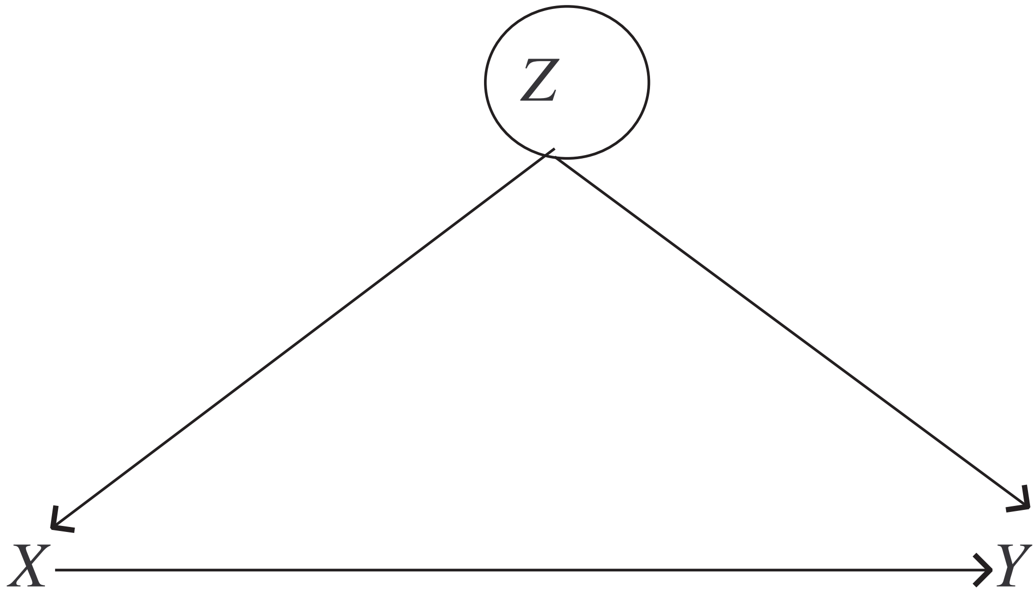
Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociological Methods and Research*, 27, 148-181.

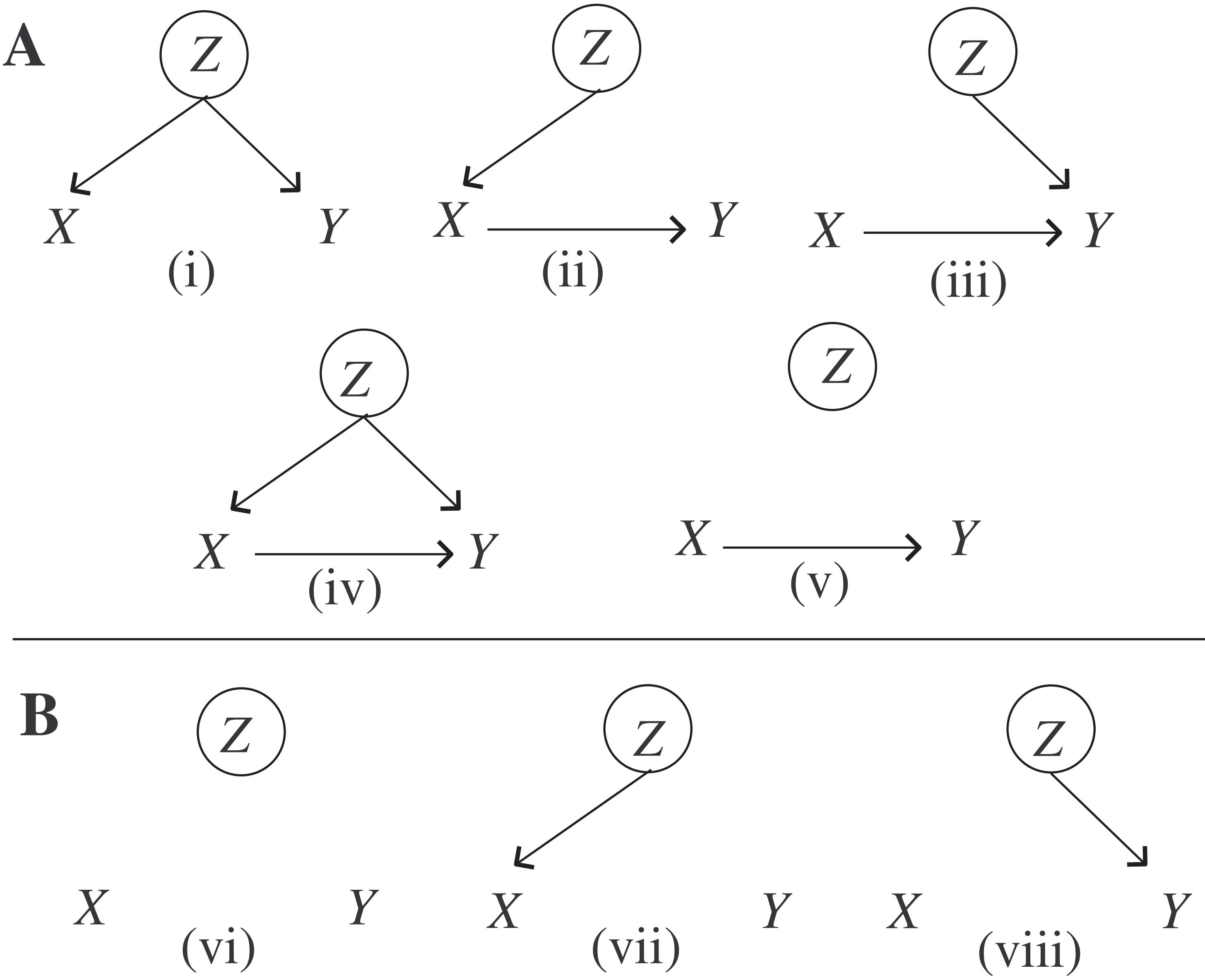
Spirtes, P., Glymour, C., and Scheines, R. (1999). On the possibility of inferring causation from association without background knowledge. In: *Computation, Causation, and Discovery.* Eds. C. Glymour and G. Cooper. Cambridge, MA: The MIT Press.

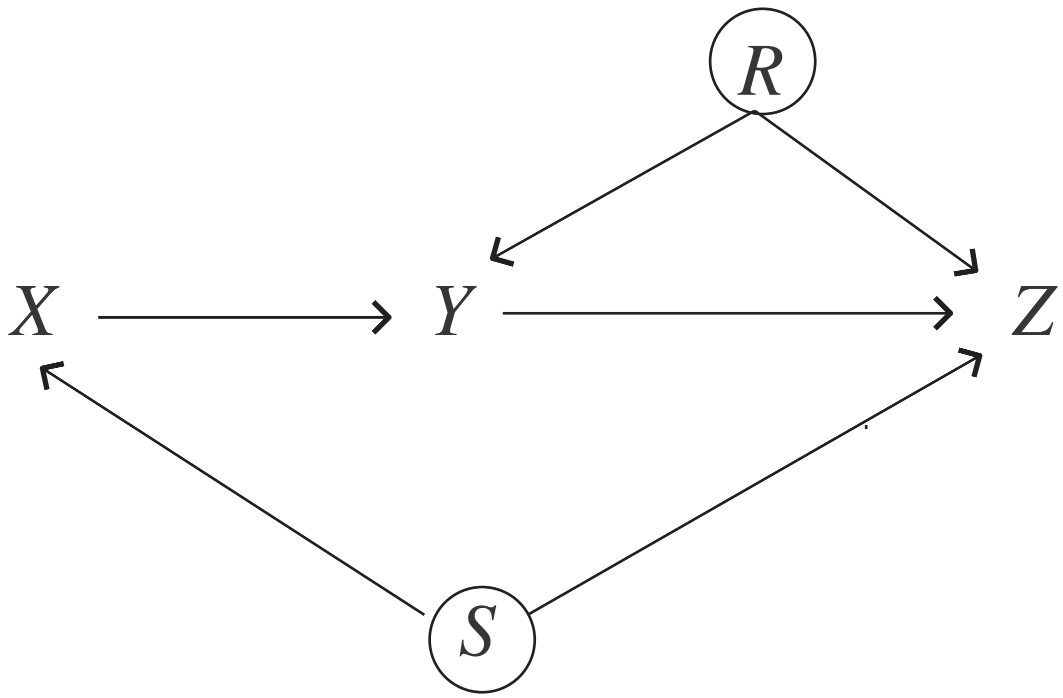
Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. Sociological Methods and Research, 27, 148-181.

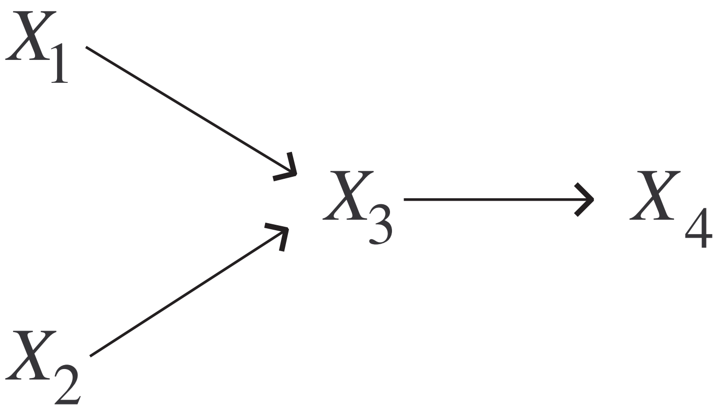
Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *Journal of the American Statistical Association*, 75, 963-972.

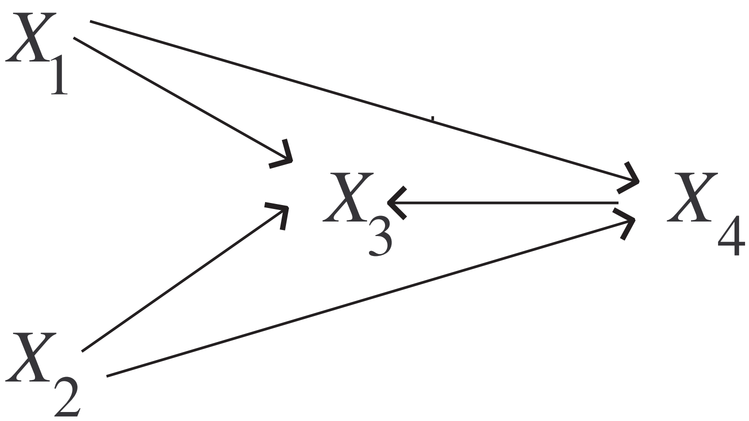
Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5, 161-215.


Z
X
Y


A
Z
X
Y
(i)
(ii)
(iii)
(iv)
(v)
B
(vi)
(vii)
(viii)


R
X
Y
Z
S


X_1
X_2
X_3
X_4


X_1
X_2
X_3
X_4