

In: **Computation, Causation, and Discovery**. Eds. P Glymour and G. Cooper.
Menlo Park, CA, Cambridge, MA: AAAI Press / The MIT Press. 1999. pp. 305-321.

CHAPTER EIGHT

On the Impossibility of Inferring Causation from Association without Background Knowledge

James M. Robins and Larry Wasserman

Spirtes, Glymour and Scheines, in their book *Causation, Prediction, and Search* (1993) and Pearl and Verma, in their paper "A Theory of Inferred Causation" (1991) make the startling claim that it is possible to infer causal relationships between two variables X and Y from associations found in observational (nonexperimental) data without substantive subject-matter-specific background knowledge. When causal relationships are represented by directed acyclic graphs (DAGs), Spirtes, Glymour, and Scheines argue that their claim follows, mathematically, from two reasonable assumptions: (1) the sample size is sufficiently large and (2) the distribution of the random variables is faithful to the causal graph. In particular, Spirtes, Glymour, and Scheines have shown that under their faithfulness assumption, there exist methods for identifying causal relationships which are asymptotically (in sample size) correct.

However, we shall show that Spirtes, Glymour, and Scheines's (1993) asymptotics implicitly assume that the probability of there being "no unmeasured common causes" of X and Y is positive and not small relative to sample size. We prove that, under an asymptotics for which the probability of "no unmeasured common causes" is small relative to sample size, causal relationships are nonidentifiable from the data alone, even when we assume distributions are faithful to the causal graph. We argue that, in observational epidemiologic, econometric, and social scientific studies, a formal asymptotic analysis that models the probability of "no unmeasured common causes" as small relative to sample size accurately reflects the beliefs of practicing professionals. We argue that these beliefs derive both from experience and from the fact that the world contains so many potential unmeasured common caus-

es (i.e., confounders) that it is a priori highly unlikely that not a single one actually causes both X and Y . We conclude that, in observational studies, small causal effects can never be either reliably ruled in or ruled out; furthermore, one should not make the leap from even relatively large empirical associations to causation without substantive subject-matter-specific background information.

1. Introduction

Several authors have used directed acyclic graphs (DAGs) as a basis for inferring causal relationships from nonexperimental (i.e., observational) data. For example, Spirtes, Glymour and Scheines (1993) and Pearl (1995) show that many issues in causal inference can be illuminated using causal DAGs. In a causal DAG, the presence or absence of an arrow between two variables represents the presence or absence of a direct causal effect. Robins (1995, 1997) shows that these DAG models are isomorphic to a causal model of Robins (1986, 1987). In Robins's (1986, 1987) approach, the causal DAG is based on background information, such as time order and subject matter knowledge of potential confounding factors. The possibility that additional unsuspected unmeasured confounders may well exist is explored through sensitivity analysis (Robins 1997) and Robins, Rotnitzky, and Scharfstein (1999). In contrast, Spirtes, Glymour, and Scheines (1993) and Pearl and Verma (1991) go further and claim it is possible to deduce aspects of the DAG from the data in the absence of background information. That is, they claim that they can "deduce causation from associations" in the data without substantive subject matter knowledge. Spirtes, Glymour, and Scheines show that their methods for doing so are correct, asymptotically in sample size, assuming a condition called "faithfulness." Statisticians, philosophers and epidemiologists have been skeptical of such claims. Humphrey and Freedman (1996) and Freedman (1993), in extended critiques, attacked many of Spirtes, Glymour, and Scheines's arguments and assumptions, including both the faithfulness assumption itself and Spirtes, Glymour, and Scheines's treatment of unobserved potential common causes. Robins (1997, section 11) centered his critique around the possibility of unobserved potential confounders (i.e. common causes). But Spirtes, Glymour, and Scheines explicitly allow for unobserved confounders.

However, we will show that the Spirtes, Glymour, and Scheines asymptotic analysis implicitly assumes that the sample size is not only large, but is large relative to the number of potential confounders. As a consequence, Spirtes, Glymour, and Scheines's asymptotics also implicitly assumes that the probability that there exists "no unmeasured common causes (confounders)" is not small relative to sample size. We investigate the implica-

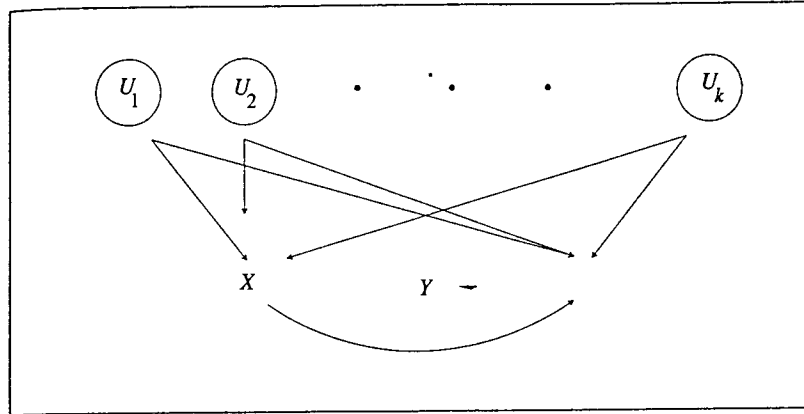


Figure 1. Directed acyclic graph for the first example.

tions of violations of this latter assumption. Specifically, we first carry out a formal analysis under an asymptotics for which the probability of “no unmeasured common causes” is small relative to sample size. We perform this analysis from a Bayesian point of view because Bayesian reasoning is a coherent and important normative approach to inference and decision making under uncertainty. The specific result we derive is that, under some mild restrictions on our prior distributions, the Bayes factor for the presence of a causal relation is not consistent; i.e., the Bayes factor remains bounded away from 0 and infinity as sample size tends to infinity when the prior probability of no unmeasured confounders is small relative to sample size. Thus a causal hypothesis is not decidable, even asymptotically, based on data alone. A simple formula emerges from our analysis that allows one to examine the sensitivity of causal inferences to the prior probability of “no unmeasured confounders.”

Finally we argue that in most epidemiologic, econometric, or social scientific studies, the beliefs of subject matter experts are accurately represented by an asymptotics that models the probability of “no unmeasured common causes” as small relative to sample size. In most observational studies, there will be measured as well as unmeasured potential confounders. Our results apply directly to this case by arguing conditionally within levels of the measured potential confounding factors.

2. A Simple Causal Model

Let (X, Y, U_1, \dots, U_k) be random variables where X and Y are observed and U_1, \dots, U_k are unobserved. X is known to occur before Y and we are inter-

ested in the causal relationship between X and Y . The random variables U_1, \dots, U_k are potential confounders. They are to represent all potential confounders in the universe, so k might be very large. To begin, assume the variables have a joint normal distribution. (We will drop this assumption in section 4, but for now it simplifies the analysis.)

For simplicity, assume $U_i \sim N(0, 1)$, $i = 1, \dots, k$ independently of each other and of X and Y . Furthermore, assume that

$$X = \sum_{i=1}^k \alpha_i U_i + \mu_x + \sigma_x \varepsilon_x \quad (1)$$

$$Y = \sum_{i=1}^k \beta_i U_i + \theta X + \mu_y + \sigma_y \varepsilon_y \quad (2)$$

where ε_x and ε_y are independent standard normals and $\Theta \alpha_i, \beta_i, \mu_x, \mu_y, \sigma_x$ and σ_y are parameters. The model is illustrated by the directed acyclic graph in figure 1. In the social science literature, our model would be referred to as a linear structural equations model (SEM) with latent variables U_i , $i = 1, \dots, k$.

Let $s = (s_1, \dots, s_k)$ denote a string of 0's and 1's and let

$$S = \{s = (s_1, \dots, s_k); s_i \in \{0, 1\}, i = 1, \dots, k\}.$$

Let G_s denote the subgraph which contains only those confounders such that $s_i = 1$. Let G_s^* be the graph G_s with the arrow from X to Y removed. For example, $G_{(1,0,\dots,0)}$ is the graph in which the only confounder is U_1 . If the data were actually generated by graph $G_{(1,0,\dots,0)}$ we say that, of the potential confounders (U_1, \dots, U_k) , only U_1 is a (true) confounder.

Remark: Note that a graph G_s having an arrow from X to Y corresponds to $\theta \neq 0$ in equation (2). The graphs G_s^* correspond to $\theta = 0$. Similarly, the graph $G_{(1,0,\dots,0)}$ in which the only confounder is U_1 corresponds to $\alpha_1 \beta_1 \neq 0$ and $\alpha_i \beta_i = 0$, $i = 2, \dots, k$, in the SEM model equations (1-2). This reflects the fact that U_m is a confounder for (i.e., an unmeasured common cause of) X and Y if and only if $\alpha_m \beta_m \neq 0$.

Let \mathcal{G} be all 2^k possible graphs with an arrow from X to Y ; let \mathcal{G}^* be all 2^k possible graphs without an arrow from X to Y ; and let $\mathcal{H} = \mathcal{G} \cup \mathcal{G}^*$. Finally, let ψ_s be the parameters corresponding to a given graph G_s and similarly let ψ_s^* be the parameters corresponding to a given graph G_s^* .

The prior is defined as follows. Each graph $G \in \mathcal{H}$ has prior probability $1/2^{k+1}$. Let ψ denote the parameters in a subgraph G . We assume that the prior for ψ is absolutely continuous with respect to Lebesgue measure with density $\pi(\psi)$, say. (This is equivalent to putting a single prior on the big graph which is a mixture of priors, each of which is singular and hence gives positive measure to certain confounders being absent.) Let A be the event "there is an arrow from X to Y " and let C_i be the event " U_i is a confounder." Then, a priori, (1) $P(A) = 1/2$, (2) $P(C_i) = 1/2$ and (3) the probability

$$P(C_1^c \cap \dots \cap C_k^c)$$

of no (unmeasured) confounders is 2^{-k} . It is important to note that any given confounder has positive probability of being absent. Nonetheless, the probability that there exist no confounders is small. As discussed later, this seems to capture precisely the way most subject matter experts would treat an observational study.

Spirtes, Glymour, and Scheines require one further property called “faithfulness.” This is a technical condition which asserts that, conditional on a graph G being the true graph generating the data, the underlying distribution possesses only those independences shared by all distributions whose densities can be factorized according to the graph. Informally, an unfaithfulness occurs if variables are independent, not by the absence of an arrow in the graph, but by a coincidental cancellation of parameter values. See Spirtes, Glymour, and Scheines (1993, page 35) for more details. Rather than saying that Spirtes, Glymour, and Scheines require faithfulness, a more accurate statement is this: in their asymptotic analysis, they exclude unfaithful distributions. Under any prior on the parameter space which parameterizes the set of distributions over the graph and which is absolutely continuous with respect to Lebesgue measure, the unfaithful distributions are a set of measure 0. Thus, Spirtes, Glymour, and Scheines argue, and we agree, that ignoring these distributions is harmless. Our assumption that the priors $\pi_i(\bullet)$ and $\pi_i^*(\bullet)$ are smooth densities with respect to Lebesgue measure respect Spirtes, Glymour, and Scheines’s faithfulness condition since, a priori, $P(D) = 0$ where D is the event that unfaithfulness occurs. This follows since $P(D) = \sum_G P(D|G)P(G)$ and $P(D|G) = 0$ for all $G \in \mathcal{H}$.

Remark: Note that the event that the variable U_i is not a common cause of X and Y (i.e., the event $\alpha_i\beta_i = 0$) has positive prior probability even though the event $\alpha_i\beta_i = 0$ has Lebesgue measure zero. This prior represents beliefs implicitly held by Spirtes, Glymour, and Scheines (and with which we agree) that, for any given variable U_i , the probability that it does not cause both X and Y is nonzero.

3. Analysis Using Faithfulness

Spirtes, Glymour, and Scheines use the faithfulness assumption to deduce the absence of a causal relation as follows. (Pearl and Verma use a similar assumption called “stability.”) Suppose that n is large and that, in this large sample, we discover that X and Y are independent (i.e., the sample correlation between X and Y is exactly zero). This observed independence is consistent with the graph G_0^* in which there are no confounders and there is no ar-

row from X to Y . In every other graph $G \in \mathcal{H}$, the only way to achieve independence of X and Y is to violate faithfulness. Thus, under the assumption of faithfulness and the assumption that the sample is large enough to reliably estimate lack of dependence between X and Y , we conclude that the only graph consistent with the data is G_0^* . We have deduced no causal relationship between X and Y and no confounders, simultaneously. We conclude that were Spirtes, Glymour, and Scheines's faithfulness analysis valid, we could infer no causation from no empirical association in the absence of any substantive subject-matter-specific background information and, indeed, even without knowledge of the real world variables represented by X and Y .

Remark: Of course, the event that the sample correlation between X and Y is precisely zero is an event that has probability zero when the data have been generated by any graph in \mathcal{H} including G_0^* . Therefore, in practice, Spirtes, Glymour, and Scheines suggest that when the p -value for a test of the hypothesis that "the population correlation is zero" is sufficiently large (say, greater than .5) and the sample size n is large, one concludes that X and Y are independent and thus, by faithfulness, that graph G_0^* generated the data.

4. A Formal Asymptotic Analysis

The Spirtes, Glymour, and Scheines analysis assumes the sample size n is large but says nothing about the relative size of n and the number of potential confounders k . We will now carry out a formal Bayesian statistical analysis which takes the relative size into account. We are interested in the Bayes factor

$$B_n = \frac{P(A^c | Z^n)}{P(A | Z^n)} = \frac{\sum_{s \in S^*} m_s^*}{\sum_{s \in S} m_s} \quad (3)$$

where

$$m_s = \int L(\psi_s) \pi_s(\psi_s) d\psi_s$$

and $L(\psi_s)$ is the likelihood function for G_s , $Z^n = (Z_1, \dots, Z_n)$ and $Z_i = (X_i, Y_i)$ are independent, identically distributed observations from the model. Similarly

$$m_s^* = \int L(\psi_s^*) \pi_s^*(\psi_s^*) d\psi_s^*$$

Bayes factors are discussed in Jeffreys (1961, chapter 3) and Kass and Raftery (1995). B_n has the formal interpretation as the posterior odds of the event A^c that X does not cause Y , since each graph has the same prior probability. Thus, if B_n is very large, we would infer that there is no causal rela-

tion. If B_n were near zero, we would infer a causal relation. The observables (X, Y) are bivariate normal and

$$T = \left(\sum_i X_i, \sum_i Y_i, \sum_i X_i^2, \sum_i Y_i^2, \sum_i X_i Y_i \right)$$

is a five dimensional minimal sufficient statistic. Apart from G_0 and G_0^* , each graph contains more than five parameters. Graph G_0 contains exactly five parameters, i.e., $E(X)$, $E(Y)$, $\text{var}(X)$, $\text{var}(Y)$, $E(XY)$. Graph G_0^* contains only four parameters, since $E(XY) - E(X)E(Y)$ is zero.

Consider any graph except G_0^* . Reparameterize ψ as (v, τ) where $v = h(\psi)$ is a five dimensional identified parameter and τ is such that reparameterization is smooth and 1-1. This leaves τ unidentified. Such a reparameterization is possible since we are dealing with an exponential family. One possible choice of v is $v = (E(X), E(Y), E(X^2), E(Y^2), E(XY))$. Because (X, Y) is bivariate normal, the likelihood is a function of v , $L(v, \tau) = L(v)$. For simplicity, in this section, assume that the marginal prior for v is the same in each subgraph. A more realistic analysis which does not make this assumption is provided in section 6. For any subgraph except G_0^* , we have

$$\begin{aligned} m &= \iint L(v, \tau) \pi(v, \tau) dv d\tau \\ &= \iint L(v) \pi(v, \tau) dv d\tau \\ &= \iint L(v) \pi(v) dv \\ &= c_n, \text{ say.} \end{aligned}$$

However, m_0^* will typically not equal c_n since distributions for graph G_0^* have only four free parameters. It follows from equation 3 that

$$B_n = \frac{(2^k - 1)c_n + m_0^*}{2^k c_n} = 1 - 2^{-k} + B^* 2^{-k} \quad (4)$$

where $B^* = m_0^*/c_n$. Note that B^* is precisely the Bayes factor for comparing G_0^* versus G_0 , i.e., for testing the presence or absence of an arrow from X to Y under the assumption of no confounders.

Now consider the limiting behavior of B_n . For this analysis we rely on well known results about the asymptotic behavior of likelihoods and integrated likelihoods (Kass, Tierney and Kadane 1990; Kass and Wasserman 1995; Haughton 1988). Generally, the behavior of Bayes factors may be summarized as follows; (details can be found in section 9). In comparing two models M_1 and M_2 , where M_1 is nested in M_2 , the following happens. If the true density p is such that $p \in M_2 \cap M_1^c$ then B_n tends to 0 exponentially quickly almost surely. If $p \in M_2 \cap M_1$ then B_n tends to infinity at rate $n^{d/2}$ where d is the difference in the dimensions of the two models. The latter case is an instance of Occam's razor in which the Bayes factor chooses the simpler model when both contain the true distribution.

Turning to our case, and applying the above reasoning to B^* , we see the following. Suppose the true model is any model other than G_0^* . Then B^* tends to 0 almost surely and hence $B_n \rightarrow 1-2^{-k} \approx 1$. Thus, the posterior odds that X causes Y converges to the prior odds and thus the causal hypothesis cannot be decided.

If G_0^* were true—and this is the important case for it corresponds to the case in section 3—then B^* tends to infinity at $\sqrt{(n)}$, i.e., $1/B^* = O_p(n^{-1/2})$. Indeed, as shown in section 9, even the maximum of B^* over all data configurations (which occurs when the sample correlation between X and Y is zero) only tends to infinity at rate $\sqrt{(n)}$.

Now, if k were fixed and n were allowed to grow, we could vindicate the Spirtes, Glymour, and Scheines faithfulness analysis that X does not cause Y since then B_n would tend to infinity in probability. But, even with n large, we might be concerned that k is large. Mathematically, we can capture this by allowing $k = k_n$ to grow with n . We do not mean that, literally, the number of confounders grows with sample size. Rather, we mean that in any asymptotic analysis we must ensure that we account for the fact k and n can simultaneously both be large. Spirtes, Glymour, and Scheines implicitly assume that $k_n = o(\log n)$. However, from equation 4 we see that if k_n is such that $k_n - (\log n)/(2 \log 2) \rightarrow \infty$ as n goes to infinity, then $B_n \rightarrow 1$ (in probability). We have arrived at the following result.

THEOREM 1. If $k_n - (\log n)/(2 \log 2) \rightarrow \infty$ as $n \rightarrow \infty$ then, whatever model is true, $B_n \rightarrow 1$ in probability.

In words, if the number of confounders is large relative to the log sample size, the data alone cannot inform us about the causal hypothesis. In the model studied in this section, as the number of potential confounders k increases, the prior probability $1/2^k$ of no unmeasured confounders decreases. Thus we can recast theorem 1 as saying that when the prior probability of no unmeasured confounders is $o(n^{-1/2})$, the data alone will not allow us to infer that X does not cause Y . In section 6 we show, in a much more general set-up, that magnitude of the prior probability of “no unmeasured confounders” relative to sample size is the crucial determinant of whether one can infer that X does not cause Y .

4.1 Spirtes, Glymour, and Scheines Faithfulness Analysis Revisited

When the prior probability of no unmeasured confounders is $o(n^{-1/2})$, and X and Y are uncorrelated in the data, the Spirtes, Glymour, and Scheines faithfulness analysis of section 3 leads to the inappropriate conclusion that, with near certainty, there is no arrow from X to Y . The error committed by the faithfulness analysis is to conclude that X and Y are truly independent if the

sample correlation is zero and the sample size is large. To see why, it is convenient to first consider the case in which the prior probability of no unmeasured confounders is exactly zero as in Robins (1997). That is, we are certain that graphs G_0 and G_0^* do not generate the data. The appropriate inference to draw from the sample correlation between X and Y being zero is that, with high posterior probability, the true correlation coefficient between X and Y lies in a small interval around zero. Since, by faithfulness, among all the graphs in \mathcal{H} only G_0^* is consistent with the correlation coefficient of zero, we further conclude that the true correlation lies in a small interval around zero, but (with posterior probability 1) is not zero itself. Thus, a posteriori, we would infer there exists a true nonzero correlation between X and Y . Since this correlation can be explained by the unmeasured confounders, whether or not there is an additional arrow from X to Y , we remain uncertain as to whether an arrow from X to Y indeed exists.

More specifically, if we modify our linear SEM example by setting the prior probability of the two graphs G_0 and G_0^* with no unmeasured confounders equal to zero and increase the prior probability of each of the other graphs to $1/(2^{k+1} - 2)$, we find that the posterior probability that X causes Y is exactly the prior probability (i.e., $B_n = 1$) for all data realizations without resorting to asymptotics.

Now suppose the prior probability that there are no unmeasured confounders is small relative to sample size, although nonzero. Then, with high posterior probability, the true correlation between X and Y will now lie in a small interval around zero that includes zero itself. However, we can infer that X does not cause Y only if our posterior probability is nearly one for the event that the true value of the correlation is zero and thus, by faithfulness, that graph G_0^* is the true graph. But, since the posterior probability of a zero true correlation is zero when, with certainty, there are no unmeasured confounders, this probability will not jump discontinuously to near 1 as we begin to increase (from zero) our prior probability of “no unmeasured confounders.”

5. A More Complex Example

In the example in the previous section, Spirtes, Glymour, and Scheines’s faithfulness analysis led to an inappropriate conclusion of the absence of a causal relationship. We now show that the reverse case can occur. Specifically, we show Spirtes, Glymour, and Scheines’s faithfulness analysis can lead to an inappropriate conclusion that a causal relationship exists.

Consider the following “faithfulness analysis” discussed by Pearl and Verma (1991) and Spirtes, Glymour, and Scheines (1993). Let X , Y , and Z be

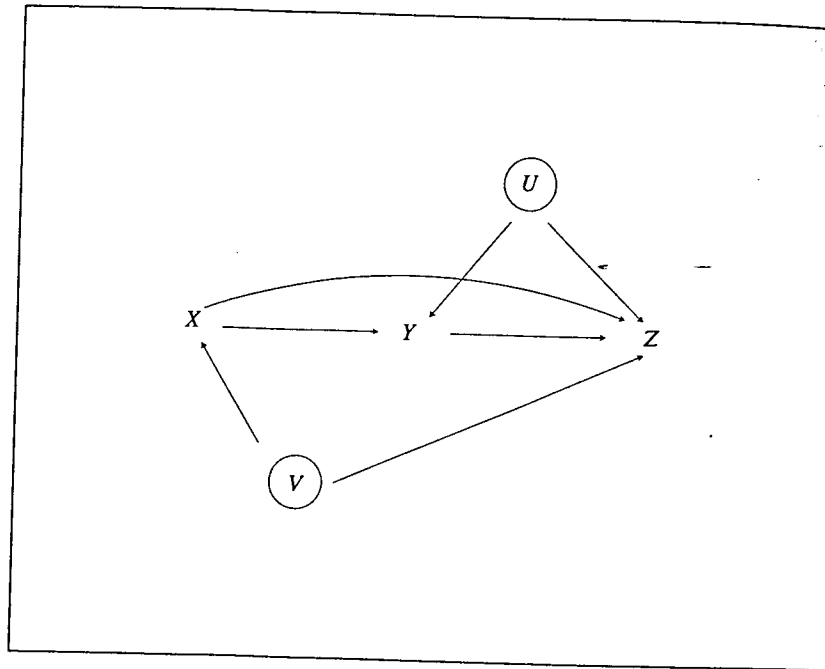


Figure 2. Directed acyclic graph for the second example.

time ordered and, let U represent all potential confounders between Y and Z and let V represent all potential confounders between X and Z ; see figure 2. For compactness of notation, in figure 2, we have not split up these potential confounders into many different potential confounders as we did before. The question of interest is whether Y is a direct cause of Z .

Suppose we observe a very large i.i.d. sample $\{(X_i, Y_i, Z_i); i = 1, \dots, n\}$ from a trivariate normal. In the sample we find the following facts (with p -values in parentheses):

1. X and Z are dependent ($p < 10^{-6}$);
2. Y and Z are dependent ($p < 10^{-6}$);
3. X and Z are independent given Y ($p = 1.000$).

Suppose now, following Spirtes, Glymour, and Scheines, we assume that facts 1 to 3 are also true in the population and we impose faithfulness. Then from fact 3 and the assumption of faithfulness, we deduce that the arrow from X to Z can be removed. We can now prove, by contradiction, that Y causes Z . Specifically, we shall prove that, if there were no arrow from Y to Z , and facts 1 to 3 held in the population, then the distribution of the data would be unfaithful to the graph in figure 2, which is not allowed.

Suppose then there is no arrow from Y to Z . If we remove U from the graph but leave V , we would deduce a dependence between X and Z given Y contradicting fact 3; thus U cannot be removed. Likewise, if we remove V from the graph but leave U , we would deduce a dependence between X and Z given Y contradicting fact 3; thus V cannot be removed. (These facts follow from well known properties of the multivariate normal distribution or, more simply, from the d-separation properties of DAGs; see Spirtes, Glymour, and Scheines [1993 page 36].) What happens if we leave both U and V ? In that case, fact 3 cannot hold unless the parameters of the distribution of the graph cancel fortuitously to give fact 3. But this is precisely a violation of faithfulness and is not allowed. Hence, we must remove U and V from the graph. But with U and V removed, Y and Z are independent, violating fact 2. Thus, following Pearl and Verma (1991) and Spirtes, Glymour, and Scheines, we conclude that there must be an arrow from Y to Z (and furthermore, by faithfulness, that V and W are absent). We have deduced a causal relationship between Y and Z . Note that the logic leading to the deduction of a causal relationship between Y and Z continues to hold unchanged even if the sample correlation $\hat{\rho}$ between Y and Z is only 10^{-4} .

If we now expand U and V to show that each explicitly represents many potential confounders, then the Bayesian reasoning of section 4 can be used here too to show that (1) the analysis is undecidable (in the sense that the posterior odds that Y causes Z do not converge to zero or to infinity as $n \rightarrow \infty$) if the number of potential confounders is allowed to be large relative to n and thus the prior probability of “no unmeasured confounders” is small relative to sample size, but (2) if the number of potential confounders is small relative to n , then the Spirtes, Glymour, and Scheines faithful analysis is vindicated and we could conclude that Y causes Z even if the sample correlation $\hat{\rho}$ was but 10^{-4} . Again, when the prior probability of “no unmeasured confounders” is small relative to sample size, the error committed by the faithfulness analysis is to conclude that facts 1 to 3 hold in the population solely because they hold in the sample and the sample size is large.

6. Robustness to the Assumptions

In section 4, we have shown that, when the probability of there being “no unmeasured confounders” is small relative to sample size, the Spirtes, Glymour, and Scheines faithful analysis can lead to the unwarranted conclusion that X does not cause Y under our model. However, our model relied on three simplifying assumptions. These assumptions were (1) normality, (2) the same prior was used for the parameters v in each graph and (3) each subgraph was given equal prior probability; none of them are substantively en-

tirely plausible. (As an example, suppose one believes that if X caused Y ($\theta \neq 0$), then most likely $\theta > 0$. In that case, one might expect that, in violation of (2), the prior probability that $\text{cov}(X, Y) > 0$ would be greater under model G_s with $\theta \neq 0$ than under model G_s^* with $\theta = 0$.) We will now show that dropping assumptions 1 to 3 does not materially change our conclusions.

Consider any fixed subgraph G . Let U denote all the unobserved random variables in this graph. Write the joint density as

$$f^G(x, y, u) = f_1^G(x, y) f_2^G(u | x, y)$$

We can think of the densities f_1 and f_2 themselves as parameters; this includes parametric and nonparametric approaches simultaneously. The pair (f_1, f_2) lie in a space of densities \mathcal{F} equipped with an appropriate σ -field \mathcal{B} . Let π^G be the prior on this space. Now π^G induces a prior $\bar{\pi}^G$ on f_1 . Our previous assumption was that $\bar{\pi}^G$ did not depend on G . Instead, we shall make a weaker assumption. Recall that \mathcal{G} is the set of all subgraphs in which there is an arrow from X to Y and \mathcal{G}^* is the set of all subgraphs in which there is not an arrow from X to Y . For convenience, let γ denote the prior for f_1 in the graph G_0 .

Assumption. With the possible exception of the prior for the graph G_0^* , the priors $\bar{\pi}^G$ are mutually absolutely continuous with respect to each other. Moreover, there exist constants, $0 < b < B < \infty$, independent of k , such that for each $G \in \mathcal{H} - \{G_0^*\}$

$$b < \text{ess inf}_{f_1} \left| \frac{d\bar{\pi}^G(f_1)}{d\gamma(f_1)} \right| < \text{ess sup}_{f_1} \left| \frac{d\bar{\pi}^G(f_1)}{d\gamma(f_1)} \right| < B \quad (5)$$

where the essential infimum and supremum are with respect to γ .

The assumption merely asserts that the prior for f_1 over the various subgraphs cannot vary wildly. The reason for excluding G_0^* is that this graph imposes an independence not found in the other graphs and we can imagine using, therefore, a prior on a lower dimensional submanifold.

Another assumption we made was that each subgraph had equal prior probability. Instead, let each graph have nonzero prior probability p_G and let p_0^* denote the prior probability of G_0^* . Let $P(A) = \sum_{G \in \mathcal{G}} p_G$ be the prior probability of an arrow from X to Y .

Now the posterior odds that X does not cause Y is

$$B_n = \frac{\sum_{G \in \mathcal{G}^*} p_G m_G}{\sum_{G \in \mathcal{G}} p_G m_G} \quad (6)$$

For any graph except G_0^* we have

$$\begin{aligned}
m_G &= \int \prod_i f_i(X_i, Y_i) d\pi^G(f_1, f_2) \\
&= \int \prod_i f_i(X_i, Y_i) d\bar{\pi}^G(f_1) \\
&= \int \prod_i f_i(X_i, Y_i) \frac{d\bar{\pi}^G}{d\gamma}(f_1) d\gamma(f_1). \quad -
\end{aligned}$$

By the assumption, it follows that $bc_n < m < Bc_n$, where c_n is m for the graph G_0 . Using similar calculations as in section 4, we have the following result.

THEOREM 2. Under the given assumptions we have

$$c_1 + c_2 B^* \leq B_n \leq C_1 + C_2 B^*$$

where

$$\begin{aligned}
c_1 &= \frac{b(P(A^c) - p_0^*)}{BP(A)}, \quad C_1 = \frac{B(P(A^c) - p_0^*)}{bP(A)}, \\
c_2 &= \frac{p_0^*}{bP(A)}, \quad C_2 = \frac{p_0^*}{BP(A)}
\end{aligned}$$

and B^* is the Bayes factor for G_0^* versus G_0 .

Without being more specific about the priors it is difficult to make precise statements about the asymptotic behavior of B^* . However, for every smooth parametric model, the best one can hope for is that B^* will increase at rate n^r for some r (usually $r = 1/2$); similar or slower behavior would be typical of a nonparametric model. Thus, as long as p_0^* is small relative to sample size (it was 2^{-k} in section 4), we are led again to the conclusion that B_n is asymptotically bounded away from 0 and infinity.

Specifically, we have the following result. Let N be the event that "there are no unmeasured confounders."

THEOREM 3. The posterior odds B_n that X does not cause Y is always bounded away from 0. Furthermore, if $B^* = O_p(n^r)$ and if $P(N) = o(n^{-r})$, then $B_n \rightarrow \psi$ in probability, where $0 < \psi < \infty$ and ψ depends on the prior odds that X does not cause Y .

We believe that our assumptions can be weakened further though we do not pursue the details here.

Remark: Theorem 2 notwithstanding, for certain priors, the data can greatly increase one's belief that X causes Y . For example, suppose (1) there was a moderately strong empirical correlation between X and Y , say $\hat{\rho} = .3$, (2) the sample size n was very large, (3) the SEM model of equations (1)-(2) was true, and (4) one believed, based on substantive background knowledge, that (a) the prior probability that X caused Y was $.5$, (b) the net magnitude of confounding by the unmeasured factors U was small

so that, if X did not cause Y ($\theta = 0$), the prior probability that the covariance ρ between X and Y exceeded .1 was quite small, and (c) there was a nonnegligible prior probability that the absolute value of θ was sufficiently large for ρ to exceed .3 even in the absence of unmeasured confounders. Then the posterior odds B_n that X does not cause Y would be small (although greater than c_1), and thus the data would have been quite informative as to the hypothesis that X causes Y .

7. Prior Beliefs

The authors and every other epidemiologist and statistician we know believe that, given any two variables X and Y , there almost always exist unmeasured confounders (i.e., common causes) linking the two variables. For example, early in their first epidemiology course, students are taught that no matter how extreme the p -value for a test of association, observational epidemiology is unable to either reliably rule in or rule out small causal effects. Only randomized trials can reliably detect such effects. These teachings imply that the subjective beliefs of epidemiologists hold that the prior probability that there are no unmeasured confounders is extremely small, if not zero (Robins, 1997, section 11). Otherwise, if sample size n were very large, then (a) as noted in the setting of section 4, when X and Y are uncorrelated in the data, we could reliably conclude that X does not cause Y and thus rule out even small causal effects, and (b) as noted in the setting of section 5, if (1) tests of the hypotheses that Z and Y and X and Z were independent, were rejected with extreme p -values (say, $p < 10^{-6}$), (2) the sample partial correlation between X and Z given Y was zero, and (3) the magnitude of the empirical correlation between Z and Y was small (say, $\hat{\rho} = 10^{-4}$), we could reliably conclude that Y has a small causal effect on Z .

It is our guess that these subjective beliefs of epidemiologists derive chiefly from two facts, one empirical and one philosophical. Empirically, in studies with large sample sizes, one typically observes highly statistically significant associations between variables which are firmly believed, on biological grounds, not to be causally associated. Philosophically, the universe contains so many unmeasured potential common causes that it is a priori highly unlikely that not a single one is an actual common cause.

In discussions with a number of economists and social scientists, it is our impression that they too believe that the probability that there are no unmeasured confounders is small if not precisely zero.

8. Practical and Philosophical Implications

We have shown that the Spirtes, Glymour, and Scheines faithful analysis can lead to inappropriate causal conclusions when the prior probability of there being “no unmeasured confounders” is small relative to sample size. Therefore, we would caution against relying on computer programs such as Spirtes, Glymour, and Scheines’s Tetrad that use the “faithfulness” analysis of section 3 as a tool for searching epidemiologic data bases for causal associations. However, our argument against the use of the Spirtes, Glymour, and Scheines “faithfulness analysis” in analyzing epidemiologic data is not an argument against using a “faithfulness” analysis in analyzing data in simple stereotyped environments where the number of unmeasured potential confounders is known to be small. Thus in artificial intelligence programs designed to allow robots to learn from data obtained in a rather stereotyped environment, the use of a “faithfulness” analysis and thus of a Tetradlike program might be extremely useful. Furthermore, our argument is not inconsistent with Pearl and Verma’s speculation that children might naturally employ an informal version of the faithfulness analysis of section 5 to learn simple causal relationships in settings where the number of alternative explanations entertained is not large. Rather, our argument against the use of a “faithfulness analysis” pertains to observational studies with moderate expected effect sizes and large numbers of potential confounding factors (such as studies of the effect of preschool Head Start programs on later high school performance or of alcohol consumption on coronary artery disease) that require large investments in study design, data collection, quality control, and data analysis (including investment in the graduate training of epidemiologists, statisticians, economists and sociologists).

We are not claiming that inferring causal relationships empirically is impossible. Randomized studies with complete compliance are a well-known example where reliable causal inference is possible. Indeed, well-supported causal inferences based on observational data are sometimes possible by (1) adjusting for or matching on measured confounders (Rubin 1974); (2) using subject matter knowledge based on experience and biology to argue, as in the remark of section 6, that the magnitude of confounding due to unmeasured factors is likely small relative to the size of the observed association; and (3) combining information from data obtained on different populations and from different types of studies, including laboratory and animal studies. The inference that cigarette smoking causes lung cancer is perhaps the best known such example. However, observational studies cannot reliably rule in or rule out small causal effects.

9. Asymptotic Behavior of Bayes Factors

We now examine the asymptotic behavior of B_n in the case in section 4. By Laplace's method,

$$c_n = L(\hat{\nu})\pi(\hat{\nu})(2\pi)^{s/2}\{\det(\sigma)\}^{1/2}n^{-s/2}(1 + O_p(n^{-1}))$$

$$m_0^* = L(\hat{\nu}^*)\pi(\hat{\nu}^*)(2\pi)^{s/2}\{\det(\sigma^*)\}^{1/2}n^{-s/2}(1 + O_p(n^{-1}))$$

where σ is the Fisher information (for a single observation) in G_0 , σ^* is the Fisher information (for a single observation) in G_0^* , $\hat{\nu}$ is the maximum likelihood estimate in G_0 and $\hat{\nu}^*$ is the maximum likelihood estimate in G_0^* . The term m_0^*/c_n tends to infinity at rate $O_p(n^{1/2})$ if A is false and tends to 0 exponentially quickly if A is true. Since $L(\hat{\nu}^*)$ is less than or equal to $L(\hat{\nu})$ with equality if and only if the sample correlation is zero, m_0^*/c_n is maximized for data sets with zero sample correlation. Even for such data sets, m_0^*/c_n tends to infinity only at rate $O_p(n^{1/2})$.

Acknowledgments

We thank David Freedman and Judea Pearl for comments on an earlier draft of this paper. Robins's research was supported by NIH Grant AI32475. Wasserman's research was supported by NIH grant R01-CA54852 and NSF grants DMS-9303557 and DMS-9357646.

References

- Freedman, D. 1993. From Association to Causation via Regression. Technical Report, 414, Department of Statistics, University of California at Berkeley.
- Haughton, D. M. A. 1988. On the Choice of a Model to Fit Data from an Exponential Family. *The Annals of Statistics* 16(1): 342–355.
- Humphreys, P., and Freedman, D. 1996. The Grand Leap. *British Journal for the Philosophy of Science* 47(1): 113–123.
- Jeffreys, H. 1961. *Theory of Probability*. 3d ed. Oxford, U.K.: Oxford University Press.
- Kass, R. E., and Raftery, A. 1995. Bayes Factors. *Journal of the American Statistical Association* 90(430): 773–795.
- Kass, R., and Wasserman, L. 1995. A Reference Bayesian Test for Nested Hypotheses with Large Samples. *Journal of the American Statistical Association* 90(431): 928–934.
- Kass, R. E.; Tierney, L.; and Kadane, J. 1990. The Validity of Posterior Asymptotic Expansions Based on Laplace's Method. In *Bayesian and Likelihood Methods in Statistics and Econometrics*, eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, 473–488. New York: North Holland.
- Pearl, J. 1995. Causal Diagrams for Empirical Research. *Biometrika* 82(4): 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*, eds. J. A. Allen, R. Fikes, and E. Sandewall, 441–452. San Francisco, Calif.: Morgan Kaufmann Publishers.
- Robins, J. M. 1997. Causal Inference from Complex Longitudinal Data. In *Latent Variable Modeling and Applications to Causality—Lecture Notes in Statistics 20*, ed. M. Berkane, 69–177. Berlin: Springer-Verlag.
- Robins, J. M. 1995. Discussion of “Causal Diagrams for Empirical Research” by J. Pearl. *Biometrika* 82(4): 695–698.
- 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(9-12): 923–945.
- 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.; Rotnitzky, A.; and Scharfstein, D. 1999. Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, ed. E. Halloran and D. Berry. Berlin: Springer-Verlag.
- Rubin, D. B. 1974. Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies. *Journal of Educational Psychology* 66(5): 688–701.
- Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. New York: Springer-Verlag.