

In: **Computation, Causation, and Discovery**. Eds. P Glymour and G. Cooper.
Menlo Park, CA, Cambridge, MA: AAAI Press / The MIT Press. 1999. pp. 323-331.

CHAPTER NINE

On the Possibility of Inferring Causation from Association without Background Knowledge

Clark Glymour, Peter Spirtes, and Thomas Richardson

1. Introduction

The distinctive power of the Tetrad II program (Spirtes, Glymour, and Scheines 1993, Scheines et al. 1994), comes from its capacity, in certain circumstances, to determine that two variables, say X and Y , do not have a common cause and that any association between them is due to the influence of X on Y or of Y on X . When other parts of the procedure, or prior knowledge, say for example, that Y does not cause X , the search procedure concludes that X causes Y . Robins and Wasserman (1996) question the possibility of reliably making such inferences without background knowledge.

Consider the two hypotheses “ X causes Y ” and “ X does not cause Y .” Robins and Wasserman (1996) concede that priors that do not assign either of these hypotheses a zero probability will, in some circumstances, lead the Tetrad II inference procedures to give correct information with probability one in the large sample limit. But concession is not their point. They prove that with a prior probability that is adjusted with sample size, the Tetrad II procedures do not converge to the truth in the large sample limit. Specifically, they consider (1) a rule that associates a prior distribution over the hypotheses “ X causes Y ” and “ X does not cause Y ” with the number of potential confounders, and (2) a rule that associates the number of potential confounders with a sample size. With their rules, they show that in the large sample limit the Tetrad II inference procedures cannot learn that there is no

direct causal connection between X and Y , in the sense that the Bayes factor approaches neither zero nor infinity as the sample size increases without bound. (They are implicitly using a zero-one loss function that assigns complete failure to the Tetrad II inference procedures if it concludes that "X causes Y " regardless of the weakness of the causal connection.)

Robins and Wasserman's rule for associating a prior distribution over the hypotheses "X causes Y " and "X does not cause Y " with the sample size is not to be taken literally; i.e. they are not advocating that an investigator's prior does or should change with sample size. What they do claim is this: for those who believe that there are a large number of confounders of almost everything, the alternative limiting analysis in which the prior does change with sample size may be more informative about their beliefs about the behavior of the Tetrad II procedures in empirical applications to large samples. In other words, one should have so much prior confidence that a sample of thousands will be confounded that, no matter what the data, the posterior probability that X does not cause Y will be less than $1/2$.

Robins and Wasserman's main nontechnical conclusion is that under plausible and widely used priors, and under a (presumably) interesting and widely used zero-one loss function, their limiting analysis casts doubt upon whether the Tetrad II inference procedures are reliable, even at large but realistic sample sizes. The questions then become (1) how plausible such priors are upon reflection, and (2) whether such a prior and loss function are actually used.

2. The Robins-Wasserman Prior

There are a number of ways in which one could argue for a prior which assigns a low probability to the hypothesis of no confounding. One could argue that given any set of variables, the correct causal graph is always complete, because a missing edge corresponds to a causal influence of exactly zero, which has Lebesgue measure zero. However, Robins and Wasserman do not adopt this prior and instead argue that even if one is willing to grant a high prior to the correct causal graph not being complete, there is still reason to adopt a prior which assigns a very low probability to the hypothesis of no confounding.

Robins and Wasserman point out that, given their prior, in the large sample limit the probability of the hypothesis that A is a cause of B is bounded away from zero, regardless of whether that hypothesis is true, or whether the contrary hypothesis that A does not cause B is true. That is correct, but is somewhat irrelevant to the conclusions drawn from the Tetrad II procedures in the case Robins and Wasserman consider, involving only two variables. The

Tetrad II procedures would never infer, from two variables only, that A is a cause of B . It would either infer that A and B were not causally connected (i.e. given the time order, that A did not cause B and there were no unmeasured confounders of A and B), or it would infer that A and B are causally connected: that is, either one influences the other *or* there is a third common cause of both A and B . If the latter hypothesis is true, then the Bayes factor for the two hypotheses approaches zero exponentially quickly almost surely, and the Tetrad II procedures succeed. In the rest of this section, we will concentrate on the case where A and B are not causally connected.

The details of the Robins and Wasserman prior are obtained by a counting argument. Given two variables A and B whose time order is known, with k potential, distinct confounders, there are $2^k - 1$ distinct graphs with confounders when A causes B , and $2^k - 1$ distinct graphs with confounders when A does not cause B . In contrast, there is just one graph without confounders when A causes B , and one graph without confounders when A does not cause B . Hence Robins and Wasserman assign a probability of $1/2^{k+1}$ to the hypothesis of no confounding, and a probability of $1 - 1/2^{k+1}$ to the hypothesis of confounding. Thus even if the probability of any particular model of confounding is low, the probability of some confounding is very high if k is large. Robins and Wasserman first consider the case where the variables are normally distributed, each graph has the same prior over the identified parameters (except for the graph where A and B are causally unconnected), and each graph is given the same prior probability. None of these assumptions is essential to their analysis, but because they make the analysis easier, this is the case we will discuss. The arguments we will give are also relevant to the case where the assumptions are dropped. Given these assumptions, they prove the following theorem, where k_n is the number of potential confounders associated with sample size n , and B_n is the Bayes factor of the no confounding model versus the confounding model at sample size n :

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.

The Robins-Wassermann prior has a number of counter-intuitive features. The two most important concerns are how it individuates confounders, and how it individuates models. In an asymptotic analysis, as $k_n - \log(n)/(2 \log 2) \rightarrow \infty$, it is clear that one cannot ordinarily have an actual list of potential confounders; all that one has is a number of potential confounders. According to the way Robins and Wasserman count graphs, there are k_n graphs with one actual confounder. If one had a list of actual confounders, this would be the right way to individuate the graphs they are considering. For example, if genetics is a possible confounder of smoking and lung cancer, and atmospheric pollution is a possible confounder of smoking and lung cancer, then the hypothesis that genetics is the only confounder of smoking and lung cancer is clearly different

from the hypothesis that atmospheric pollution is the only confounder of smoking and lung cancer. These are two empirically different hypotheses making two quite distinct claims. But if one is given only the number of possible confounders (2 in this instance) and two names, U_1 and U_2 , which are not attached to particular random variables, then in what sense is the hypothesis that U_1 is the only confounder of lung cancer and smoking different from the hypothesis that U_2 is the only confounder of lung cancer and smoking? The two graphs associated with these hypotheses are just relabelings of each other, and the relabelings don't change the hypothesis since the names were not associated with particular random variables to begin with. The theory should simply be interpreted as asserting "There is one actual confounder of smoking and lung cancer." Given this interpretation, there are not k_n different graphs with one actual confounder, but there is one graph with one potential confounder. If graphs are counted in this way, then for k_n potential confounders, there are k_n+1 different graphs, each receiving probability $1/(k_n+1)$. If this is the case, then in order for B_n not to approach infinity as n approaches infinity, $k_n - (n-1)^{1/2} \rightarrow \infty$ as $n \rightarrow \infty$.

In addition to the problem of individuating different models with the same number of potential confounders, there is the problem of individuating and counting confounders, when the confounders are unspecified. In terms of the probability distributions that can be represented, any distribution that can be represented by a graph with n latent binary variables can also be represented by a graph with one latent variable with 2^n states. In a linear model, any distribution that can be represented with n confounders of a pair of variables can also be represented by a single confounder of the pair of variables. It is not clear why, when the variables are unspecified, the larger count of potential confounders is correct, as the Robins-Wasserman analysis requires.

Another feature of the Robins-Wasserman prior that can be questioned is that (as they point out) it assigns a probability of $1/2$ to each potential confounder actually being a confounder. If the probability of each potential confounder actually being a confounder is $(1/2)^m$ for some integer m greater than 1, then even if one used their count of the number of graphs with potential confounders, k_n would have to grow more quickly than $\log(n)/(2\log 2)$ in order for B_n not to approach infinity.

The Robins-Wasserman prior also assumes that whether each potential confounder is an actual confounder is independent of whether other potential confounders are actually confounders. A reason for doubting this is illustrated in figure 1. Consider the two graphs in figure 1, where U_1 and U_2 are confounders of X and Y . In figure 1a, U_2 is a direct cause of X and Y . In figure 1b, U_2 is a direct cause of Y , but is only an indirect cause of X (via U_1 .) In the former case we say that U_2 is a *direct confounder* of X and Y (relative to U_1, U_2), and in the latter case we say that U_2 is an *indirect con-*

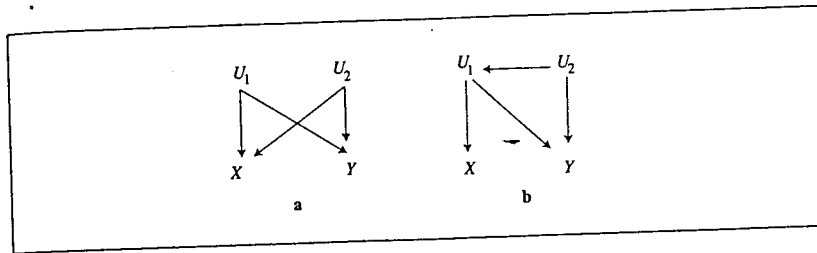


Figure 1. A reason to doubt the Robins-Wasserman prior assumption.

founder of X and Y (relative to U_1, U_2).

Note that when U_2 is an indirect confounder of X and Y , it can be a confounder of X and Y only when some other variable is also a confounder of X and Y . Now, suppose that there are k possible confounders of X and Y , but U_1 is the only possible direct confounder of X and Y . In that case, if U_1 is not a confounder of X and Y , then there are no confounders of X and Y . It follows that if each graph is equally probable, then the probability of no confounding is $1/2$. So to get the Robins-Wasserman prior it is not enough to suppose that there are k possible confounders; one must make the stronger assumption that there are k possible *direct* confounders.

Finally, note that Robins and Wasserman implicitly assume that for a system of m measured variables, the probability that U_i is a confounder of A_1 and A_2 is independent of whether U_j is a confounder of A_3 and A_4 . This cannot be the case, however, if each measured variable is a potential confounder of each other pair of measured variables. Assuming acyclicity, it is not possible for A_3 to be an actual confounder of A_1 and A_2 , and for A_2 to be an actual confounder of A_1 and A_3 . For example, if A_1, A_2 , and A_3 are all measured, and for each pair of variables A_i and A_j the only potential confounder is A_k ($k \neq i, k \neq j$), then each pair of variables has a potential confounder. However, only one pair of variables can be confounded at a time.

3. Are the Robins-Wasserman Prior and Loss Function Actually Used?

Does anyone actually hold the combination of priors and loss function that Robins and Wasserman use in their analysis? This is a somewhat difficult question to answer, because almost always no formal Bayesian analysis of a problem in the social sciences or in epidemiology is given. First, using the Robins and Wasserman rules for associating priors with the number of possible confounders, note the following priors associated with the following numbers of potential confounders: if there are 10 potential confounders of X and Y ,

the prior probability of there being no confounders between X and Y is .00098; with 20 it is .00000095; with 30 it is .0000000093; and with 40 it is .000000000091. That is, given 40 potential confounders, their prior assigns less than one chance in a trillion that there is no confounding of X and Y .

Robins and Wasserman argue that the behavior of epidemiologists indicates that they believe that the prior probability of no confounding is quite low. Epidemiologists do not accept that at large sample sizes, if X and Y are uncorrelated that we can reliably conclude that X does not cause Y and rule out even small causal effects. But of course at large sample sizes, the sample correlation is never exactly zero. If the sample correlation is not exactly zero, then one would have to do a statistical test to determine if the effect is zero. But it is well known that at very large sample sizes the results of statistical tests are extremely sensitive to a host of factors, including minor violations of the distributional assumptions, outliers, rounding error, measurement errors, etc. Moreover, no matter how large the sample size, the power of a statistical test against an alternative that is sufficiently close to the null hypothesis is low. Thus there is good reason to question the results of these tests even if one grants that the prior probability of no confounding is not tiny, especially if the alternative hypothesis is that the causal effect is very small.

Robins and Wasserman offer a second argument that epidemiologists assign a tiny prior probability to the hypothesis of no confounding. Consider the following possible evidence one might gather: the hypotheses that Z and Y , and X and Z are independent are rejected with extreme p -values (say $p < 10^{-6}$), the sample partial correlation between X and Z given Y is zero, and the magnitude of the empirical correlation between Z and Y is small (e.g. on the order of 10^{-4} .) Robins and Wasserman point out that epidemiologists would not accept this as conclusive evidence that Y has a small causal effect on Z . They conclude that this shows that epidemiologists assign a tiny prior probability to the hypothesis of no confounding. But the same arguments we made in the previous paragraph also apply here. In addition, the test of whether or not the direct effect of X on Z is zero in this case becomes a test of whether the correlation between X and Z given Y is zero. If the standardized linear coefficient of the direct effect of X on Z is r , and the correlation between Z and Y is small, the correlation between X and Z given Y is less than r . Hence, the power of a test that the direct causal effect of X on Z is zero against the alternative hypothesis that the direct causal effect of X on Z is small, is quite low.

We believe that, contrary to Robins and Wasserman's claims, the combination of priors and loss function that Robins and Wasserman use in their analysis is not compatible with the inferences that epidemiologists and social scientists are actually willing to make.

For example, according to Robins and Wasserman, inferring the presence or absence of causal relationships from nonexperimental data is possible, us-

ing the technique of adjusting for or matching on measured confounders (suggested by Rubin [1974], among others), or combining information from data obtained on different populations and from different types of studies. It is difficult to see how either of these methods could be reliable, unless Robins and Wasserman's prior is being implicitly denied.

Suppose that it is granted that the Robins-Wasserman prior accurately reflects the situation in observational studies, and that given a sample of size n , there will in general be at least \sqrt{n} -many confounding variables. At least in the case of discrete variables it would then appear impossible *in principle* to estimate the size of a treatment effect, by controlling for sufficiently many confounding variables. To simplify matters let us suppose that all variables, including confounders, are binary. In order to make any inference concerning the treatment effect, in the absence of confounders, we must have a sample of a certain size, say n_0 . However in the Robins-Wasserman world, there will then be approximately $\sqrt{n_0}$ confounders associated with this sample. This will mean that our state space increases by a factor of $2^{\sqrt{n_0}}$. Thus in order to assess the dependence between treatment and outcome, conditional upon these confounding variables, we will require a sample that is considerably larger, say of size n_1 , for otherwise our data will be too sparse for us to make a reliable inference. (If we cannot assume that the confounders are independent, then n_1 might need to be on the order of $2^{\sqrt{n_0}}$ times as large as n_0 , but this is inessential for the argument here; requiring a sample twice as large would be quite sufficient) However, once again, associated with the larger sample that we now require will be an increased number of confounders, and so on. We are caught in a vicious circle: to control for more confounders we require more data, owing to the increased state-space, but more data in turn is associated with more confounders.

Thus the method of adjusting for measured confounders is not reasonable given the Robins-Wasserman prior. Of course one could argue that one is only interested in the question of the approximate size of the causal effect (i.e. a different loss function based on how far off the predicted causal effect is from the actual causal effect), and Rubin's method is reasonable as long as the prior that there is no *strong* unmeasured confounder is near one. But an analogous defense of the Tetrad II procedures could be made.

Another technique that seems to be incompatible with the Robins-Wasserman prior, is the Wu-Hausman test of exogeneity. This test is equivalent to a test of independence, and as a test of exogeneity, is subject to failure due to violations of faithfulness (Davidson 1993). It is nevertheless widely used in econometrics.

It is less clear what prior is implicit in the example of the "smoking causes lung cancer" inference, if only because the actual method used to draw this conclusion is much less clear. Of course, laboratory and animal studies were used, but we are concerned here with the use of the statistical evidence relat-

ing human smoking to lung cancer. One type of evidence that has been marshaled for the conclusion that smoking causes lung cancer is that the association between smoking and lung cancer remains strong, even after controlling for a variety of measured possible confounders. But it is difficult to see what could justify this reasoning unless one believed that one had measured a significant proportion of the potential confounders. If there are 1000, or 10,000, or 100,000 possible confounders, but one has measured only 50 or 100 of them, it is hard to see why a strong association controlling for the measured confounders should be taken as evidence that smoking causes lung cancer. And if a strong association controlling for the measured confounders should be taken as evidence that smoking causes lung cancer, then (assuming coherence) not having a strong association controlling for the measured confounders should be taken as evidence that smoking is not a cause of lung cancer. Again, one could argue that one is only interested in the question of the approximate size of the causal effect, and the method is reasonable as long as the prior that there is no *strong* unmeasured confounder is near one; but again one could make the same argument for the Tetrad II procedures. Alternatively, one could argue that a strong association controlling for the measured confounders is not considered conclusive evidence that smoking causes lung cancer, but is instead interpreted as supportive. It is only conjunction with all of the other evidence (such as laboratory and animal studies) that the conclusion is considered conclusive. Again, however, one could make the same argument for the Tetrad II procedures. They should not be considered as conclusive evidence in favor of the hypothesis, but as one type of evidence that should be taken into account.

Robins and Wasserman remark that their result shows the implausibility of detecting very weak causal influences from nonexperimental data. We agree with their conclusion; but we think these procedures still play an important role in detecting larger causal influences. We further agree that it would be very useful to have studies that examine the sensitivity of the output of the Tetrad II procedures to different assumptions and different priors, using as a measure of success the difference between the predicted and the actual causal influences. So, in conclusion, we think Robins and Wasserman have identified an interesting research issue, but have described only implausible priors, not congruent with practice, under which automated search procedures cannot succeed.

References

- Davidson, R., and Mackinnon, J. 1993. *Estimation and Inference in Econometrics*. Oxford, U.K.: Oxford University Press.
- Epstein, R. L. 1987. *A History of Econometrics*. Amsterdam, The Netherlands: North-Holland.

- Engle, R.; Hendry, D.; and Richard, J. F. 1983. Exogeneity. *Econometrica* 51(4): 277-304.
- Hausman, J. A. 1978. Specification Tests in Econometrics. *Econometrica* 46(1): 1251-1271.
- Robins, J., and Wasserman, L. 1996. On the Possibility of Inferring Causation from Association without Background Knowledge. Technical Report, 649, Department of Statistics, Carnegie Mellon University, Pittsburgh, Penn.
- Rubin, D. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66(2): 688-701.
- Scheines, R.; Spirtes, P.; Glymour, C.; and Meek, C. 1994. *Tetrad II: Tools for Causal Modeling*. Hillsdale, N.J.: Lawrence Erlbaum.
- Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. Lecture Notes in Statistics 81. New York: Springer-Verlag.
- Wu, D.-M. 1973. Alternative Tests of Independence between Stochastic Regressors and Disturbances. *Econometrica* 41(3): 733-750.