Causality from Probability

by

Peter Spirtes, Clark Glymour, Richard Scheines

October 1989

Report No. CMU-PHIL-12

Philosophy
Methodology
Logic

Pittsburgh, Pennsylvania 15213-3890

Causality from Probability

Peter Spirtes, Clark Glymour and Richard Scheines[1]
Carnegie-Mellon University

1. Introduction

1.1 Uses of data analysis for policy generally involve causal inference

Data analysis that merely fits an empirical covariance matrix or that finds the best least squares linear estimator of a variable is not of itself a reliable guide to judgements about policy, which inevitably involve causal conclusions. The policy implications of empirical data can be completely reversed by alternative hypotheses about the causal relations of variables, and the estimates of a particular causal influence can be radically altered by changes in the assumptions made about other dependencies.[2] For these reasons, one of the common aims of empirical research in the social sciences is to determine the causal relations among a set of variables, and to estimate the relative importance of various causal factors. Even where that aim is not acknowledged it is often tacit. A question of first importance about empirical social science is therefore: how are causal relations among variables to be discovered?

1.2 The difficulty of the discovery problem

The difficulty of this question is apparent when one considers the number of possible causal models for a given set of variables. If the causal dependence of one variable on another is represented by a directed edge from a vertex representing the causal variable to a vertex representing the effect variable, then the number of possible causal structures on n variables is the number of directed graphs with n vertices, or $4^{\binom{n}{2}}$. If causal cycles are forbidden, then the number of possible causal structures on n variables is the number of acyclic directed graphs on n variables. For 12 variables the number of directed graphs is approximately 5.4×10^{39} and the

---

[1]The research in this paper was supported by the Office of Naval Research under Contract number N00114-88-K-0194. The second author was supported by a fellowship from the John Simon Guggenheim Memorial Foundation.

[2]See our discussion of the causal relations between foreign capital on political repression in [Glymour 87].

number of acyclic graphs is 521,939,651,343,829,405,020,504,063 [Harary 73]. Even when the time order of the variables is known, so that causal hypotheses in which later variables cause earlier variables can be eliminated, the number of alternatives remaining is generally very large: for 12 variables it is 7.4×10^{19}.

The social scientist who addresses a problem area where causal questions are of concern is therefore faced with an extremely difficult discovery problem, for which there are only three avenues of solution: (i) use experimental controls to eliminate most of the alternative causal structures; (ii) introduce prior knowledge to restrict the space of alternatives; and (iii) use features of the sample data to restrict the space of alternatives.

Experimental procedures for addressing social questions are much to be desired, but they are very expensive and often infeasible. Where quasi-experiments are used that control some variables but not others, the number of alternative causal structures possible *a priori* may remain very large. Generating the set of admissible causal structures from "substantive theory" is recommended routinely in methodology texts.[3] In practice publications in the social science literature usually restrict the number of alternatives considered to a very few, and the restrictions are often justified by citing prior literature or by appealing to very broad theoretical frameworks. It is anybody's guess, however, whether such appeals constitute a reliable discovery procedure. It seems at least as likely that appeals to theory introduce bias and often exclude the true causal relations among the variables of interest. What about the third avenue?

1.3 Causal inference from statistical samples

Sample data are routinely used in systematic ways for parameter estimation in a parameterized family of probability distributions, but are more rarely used explicitly or systematically to infer causal structure. To the contrary, methodologists routinely warn against such inferences. The common slogan "correlation does not imply causality " is generally given as a caveat against trying to infer anything about causal dependencies from statistical data. Methodologists routinely warn that "substantive knowledge," not sample data, should determine the causal structure of a model. Procedures that use the sample data are denounced as "data mining" or "ransacking." It would be difficult to find a textbook on statistical methodology for the social sciences that does not include these warnings.[4]

---

[3]See, for example, [Joreskog 84, Duncan 75].

[4]See [Loehlin 87], or [James et. al., 82].

For all the ferocity of the denunciation of sample based causal inference, it is hard to find any sober analysis that justifies the conviction that reliable inference of this kind is impossible. There are worst-case arguments that point out the unreliability of data based inference if the sample size is small compared to the number of variables, but these objections are readily avoided by appropriate sampling. There is the wide experience of social scientists and psychologists with a variety of "exploratory" factor analysis programs, which many people hold to be quite unreliable. But factor analytic programs involve very specific assumptions that are not in the least necessary in any possible procedure for inferring causal structure from sample data. Common factor analysis programs assume, for example, that the functional dependencies are linear, that measured variables have no direct effects on one another, and that measured variables never have effects on latent variables or factors [Loehlin 87]. Each of these assumptions may be false in a domain; none of them are essential to the idea of inferring restrictions on causal structure from sample data.

Is the complaint against sample based causal inference then simply an unfounded prejudice? If so, the best way to show as much is to provide reliable procedures for using sample data to usefully narrow the class of causal structures that are a priori possible for the data, and to prove that the procedures are reliable. That is our aim.

2. The Claims

Subject to simple and plausible principles connecting causal dependencies with statistical dependencies we will describe automatic procedures that:

> (1a) if given data for a number of random variables generated by sampling from a distribution determined by unknown causal dependencies among those variables and by an unknown probability distribution on the exogenous variables, will find a "small" set of alternative causal structures;
>
> (1b) if given the population covariances will output a set of causal structures that with probability one includes the true structure;
>
> (1c) are reliable enough to be useful on samples of realistic size;