Causality from Probability

Peter Spirtes, Clark Glymour and Richard Scheines

October 1989

Report No. CMU-LCL-89-3, CMU-PHIL-12

Laboratory for Computational Linguistics

139 Baker Hall
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213

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;

(2) provide a refiable statistical condition sufficient to warrant the conclusion that the statistical dependencies among a set of jointly distributed random variables are due in part to a random variable not contained in the set; in other words, find a reliable sufficient condition for the introduction of "latent" variables;

(3) given a causal structure, will generate all the alternative causal structures that are "statistically equivalent" to the given structure.

The procedures we will describe are partially implemented in the TETRAD II and MEASURED programs now being developed and tested at Carnegie Mellon University.

These claims are odd in that they mix causal language with talk of probability distributions. That mixture nonetheless agrees with a great deal of informal usage in biostatistics, quantitative social science and in applied statistics generally. In this paper we will provide a precise sense of causal dependency which gives a computable necessary and sufficient condition for two variables to be "directly" causally connected. This principle can be used to form an undirected graph of causal dependencies from sample data. A second computable general principle connecting causality and statistical dependency suffices to restrict the class of possible orientations of the edges of the undirected graph; i.e., gives a collection of directed graphs of causal relations from the sample data. We will describe a procedure we call MEASURED that uses these principles and that answers [1a]. We will state a theorem that for linear models ensures that the procedure satisfies the requirement of (1b), and we will describe some runs on simulated data that address (1c). Using a third principle-which is provably consistent with the first two-connecting causal structures and statistical dependencies, we will report but not prove results that give solutions to (2) and [3].

2. Causality and Statistical Dependence

Many writers [Suppes 70 ,Skyrms 80] have connected causality with statistical dependence. We suppose the following principle:

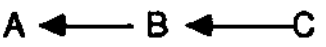
Let V be a set of random variables with a joint probability distribution. We say that variables x, y e V are **directly causally dependent** if and only if there is a causal dependency between x, y (either the value of x influences the value of y or the value of y influences the value of x or the value of some third variable not in V

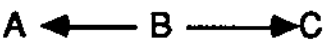
influences the values of both x and y) that does not involve any other variable in V.

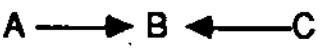
Principle I: For all x, y In V, x and y are directly causally dependent if and only If for every subset S of V not containing x or y, x and y are not statistically Independent conditional on S. [Spirtes 89b]

Principle 1 permits a procedure that generates from sample data an undirected graph whose edges represent direct causal dependencies. One needs to test for each of the conditional statistical dependencies that are possible a priori, and then apply the principle to build the graph. We have implemented this procedure in MEASURED, a part of the TETRAD II program. MEASURED tests for vanishing partial correlations of all orders. The procedure is robust. Asymptotically distribution-free tests of vanishing correlations and partial correlations are available, and the inference from the conclusion that a partial correlations does not vanish to statistical dependence is distribution free, and so therefore is any conclusions that two variables are directly causally dependent.[5]

To orient the undirected graph constructed with Principle I, a further principle is needed. Consider the following pictures of possible causal dependencies among three variables.







We say that **B Is directly causally dependent on A** provided A and B are causally dependent and the direction of causal influence is from A to B.

---

[5] The inference from vanishing partial correlations to the conclusion that two variables are *not* directly causally dependent does, however, depend on the distribution. Ironically, in the proper spirit the old slogan is reversed; correlation does imply causaGty, but absence of correlation does not imply absence of causality.

If A and B are directly causally dependent and B and C are directly causally dependent, but A and C are not, then the relations between A and B, on the one hand, and B and C on the other hand, may produce an indirect causal dependency between A and C, or they may not. Whether or not an indirect dependency is created is a function of the directions given to the connections between A and B and between B and C. If both edges are directed into B as in the lower right-hand graph, then no indirect dependency is created. If the lower right hand graph describes the causal dependencies, then information about A when added to information about B, should give further information about C, and information about C, when added to information about B, should give further information on about A. In other words, A and C should be statistically dependent conditional on any set of variables that does not contain A and C but does contain B. Moreover, the converse seems true as well, and we therefore offer

Principle II: if A and B are directly causally dependent and B and C are directly causally dependent, but A and C are not, then:

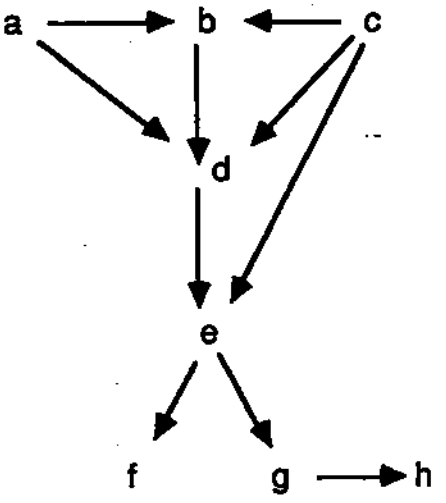
B is causally dependent on A, and B is causally dependent on C if and only if A and C are statistically dependent conditional on any set of variables containing B and not containing A or C.

The MEASURED procedure uses the partial correlations supported by a sample to obtain an undirected graph using Principle I. We plan to implement a procedure that will apply Principle II to the undirected graph and the data to obtain the admissible orientations. The result will be a set of directed graphs; the size of the set will depend on the structure of the data, but it will always be enormously smaller than the space of a priori possibilities.

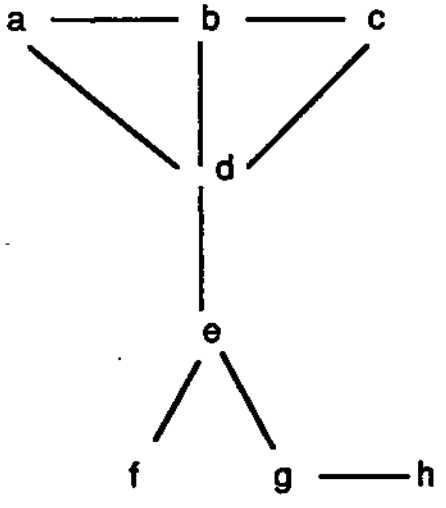
3. Examples

On large samples (n = 2000) preliminary simulation tests of the MEASURED procedures indicate that **it** is extremely reliable. The procedure that applies Principle 1 has a tendency to underfit, that is it tends to omit undirected edges corresponding to causal dependencies in the structure from which the data were obtained. These omissions can reliably be recovered by submitting the models produced by the program to an elaboration procedure that adds edges whenever possible to eliminate the impBcation of independence constraints that are not satisfied in the data.

Data for a sample of 2000 population units were generated by Monte Carlo methods from a linear model with the following causal structure (normally cfistributed variates):



The data were then given to the MEASURED procedure which produced the following undirected graph:

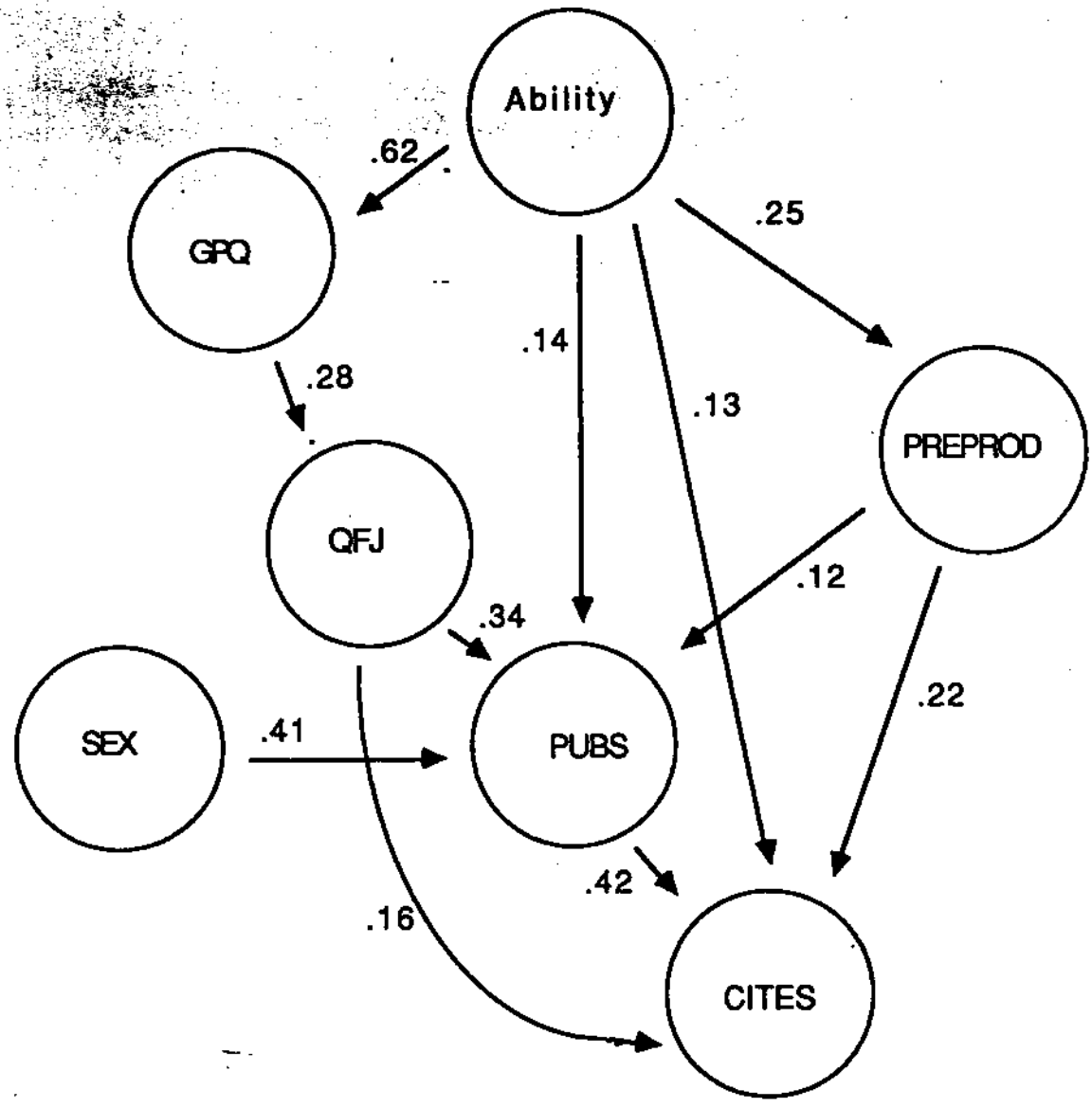


Principle II permits only one orientation of this graph, which is exactly that of the initial model used to generate the data. The omitted edge can be recovered by using elaboration techniques analogous to those of the TETRAD II program described in [Spirtes 89].

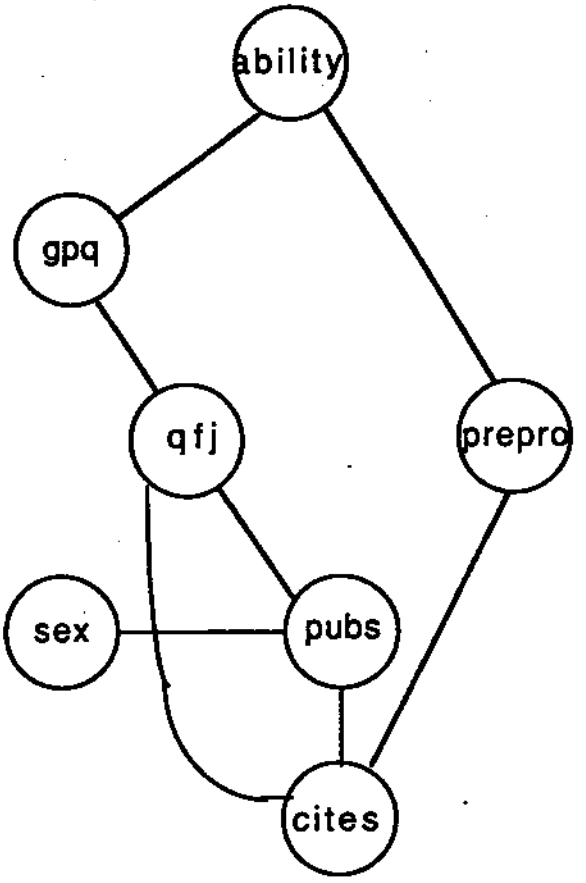
An empirical example is provided by recent work of Rogers and Maranto (Rogers 89). They studied a number of theoretical accounts of the determinants of publishing productivity in

psychology, and compared these accounts with original survey data they obtained. They first performed path analyses of six alternative models taken from the social science literature. They then formed a combined model that included all causal dependencies occurring in any of these models as well as two further dependencies. After estimating and testing the combined model, they eliminated the dependencies found not to be statistically significant. Their result is the following causal mode (with coefficients estimated assuming Gnear dependencies)[6]:

[6]GPQ is a scale formed from indicators of the quality of graduate programs; QFJ is a measure of the quality of the subject's first job; PREPRO is a measure of publications while in graduate study; PUBS a measure of publications since leaving graduate school, and CITES indicates frequency of citation of the subject's scholarly works.



If we give the correlations from Rogers and Maranto's survey data to the MEASURED procedure (whose present implementation assumes normal variates), we automatically obtain the following undirected graph *without any imposing any prior substantive restrictions:*



Output of Measured

The application of Principle I to the sample data yields eight of the eleven connections Rogers and Maranto postulate, and the three omitted edges are those which, in the linear model, have the smallest coefficients.

The time order of ability, graduate school quality, quality of first job, and publication in graduate school and other variables uniquely determines the directions that can be given to the edges in this undirected graph.

In this case the sample size is not large and we used a test for vanishing partial correlations that assumes normal variates even though some of the variables in the Rogers and Maranto study certainly are not normally distributed. The result is that we obtain many more vanishing partial correlations than seem plausible, and therefore the application of Principle II failed to determine the orientation of the edges in this case. Thanks to the time order, however, the orientation is uniquely fixed.

4. Further Connections Between Causal Dependence and Statistical Dependence

Principles I and II determine from a probability distribution a set of alternative causal structures based on the relations of statistical dependence and independence in the distribution. The principles do not generally determine a unique causal structure. Conversely, given a causal structure, the two principles determine some of the statistical dependencies and independencies in any probability distribution represented by that causal structure, but generally a causal structure does not determine all of the statistical dependencies in probability distributions that it represents. In other words, given an acyclic directed graph, Principles I and II do not tell us for all vertices x, y and all sets of vertices S, whether or not x and y are statistically independent conditional on S. There is a further principle, due to J. Pearl and his colleagues [Pearl 88], relating causal directed graphs and probability distributions that does give that information. To state it we need some definitions:

> An **undirected path** in an acyclic directed causal graph is a a sequence of vertices $v_1,...,v_n$ such that for all i between 1 and n-1, v_i and v_{i+1} are directly causally connected.
>
> A vertex v_j on an undirected path is a **collider on that path** if and only only if the graph contains the directed edges $v_{j-1} \rightarrow v_j$ and $v_{j+1} \rightarrow v_j$.
>
> A **directed path** from v_1 to v_n is an undirected path such that for all i between 1 and n-1, the graph contains the directed edges $v_i \rightarrow v_{i+1}$.
>
> A vertex y is a **descendant** of a vertex x if and only if there is a directed path from x to y.
>
> Following Pearl, we say that variables x, y are **d-separated** by set S if and only if there exists no undirected path between x and y, such that (i) every collider on that path has a descendent in S and (ii) no other vertex on the path is in S.

With these definitions we can state Pearl's principle:

> **Principle III: A directed acyclic graph represents a probability distribution on the variables that are vertices of the graph if and only if:**

for all vertices x, y, and all sets S of vertices In the graph, S d-separates x and y If and only If x and y are independent conditional on S.

Principle III is not immediately intuitive, and we do not have space to enter into issues concerning its justification, save to observe that it agrees with many intuitive principles about connections between causality and statistical dependence, and that it is satisfied by all of the usual modeling formalisms, e.g., in linear structural equation models. The connection between Principle III and Principles I and II may be unclear, however, and we report the following theorems.

Theorem 1: Let P be a probability distribution represented by an acyclic directed graph G according to Principle III. Then G Is an orientation of an undirected graph U that represents P according to Principle I.

Theorem 2: Principle III implies Principle II.

Conjecture: Let r be the set of directed graphs that represent probability distribution P according to Principle III. Then r is also the set of directed graphs obtained from P by Principles I and II.

Principle III, together with these two theorems and two other results to be stated later that use the Principle, has somejemarkable implications.

5. Characterizing Statistically Equivalent Models

There is **a** small literature on the notion of "statistically equivalent models." [Lee 87]. A few writers have attempted to give sufficient conditions for two models to be statistically equivalent. The notion is not, however, very well characterized. Intuitively, the idea is that models are statistically equivalent with respect to a set of measured variables if no statistical inference based on sample data can reliably decide between them, no matter how large the sample. In practice, however, statistical equivalence has sometimes been characterized with respect to a particular estimation procedure or test, and more often not at all.

We will say that two directed graphs sharing a common set of vertices representing measured variables are **statistically equivalent** if the graphs contain the same number of directed edges

and if, according to Principle III, the graphs imply the same collection of constraints on the measured variables.

It should be noted that for graphs in which there are no "latent" or unmeasured common causes, this condition implies that two graphs are statistically equivalent if and only if they represent exactly the same set of statistical dependencies and independencies. For graphs with latent variables this is not the case, since the constraints among the measured variables that such graphs imply (in virtue of the statistical dependencies and independencies they imply for all variables, both latent and measured) are not confined to statistical dependencies and independencies.

An immediate corollary to theorems I and II and the conjecture is that Principles I, II and III provide a more or less feasible[7] procedure for determining all of the models statistically equivalent to any given model without latent variables. The procedure takes the given graph, forms its undirected graph, and calls Principle III whenever the application of Principle II requires a fact about statistical dependence. This procedure will be implemented in the MEASURED program. The more general problem of finding a feasible procedure to characterize the class of models statistically equivalent to a given latent variable model has still to be solved.

6. Inferring Latent Variables

6.1 Latent Variables

Coupled with robust tests of vanishing partial correlations (or, for that matter, any robust test of statistical conditional independence) the procedures we have described provide a complete, almost distribution free, asymptotically reliable discovery procedure when all of the sources of statistical dependence between variables are due to causal relations among the measured variables. But what if they isn't? What if there are unmeasured factors operating? Unmeasured variables are nothing mysterious. If one goes to any sodal science data bank one will find that variables measured in one data set are not measured in another. Unmeasured variables and just that-unmeasured; they need not be (and we expect generally are not) unmeasurable. It would seem that a central problem for scientific discovery in the social sciences is to find a reliable way to

[7]The procedure is clearly worst case exponential, and we guess that the problem of deciding statistical equivalence is NP hard. By "feasible" we mean a procedure that will work in reasonable time on available computers on models of the size one typically finds in the social and behavioral sciences (exclusive of econometrics).

decide whether or not the statistical dependencies found among some of the variables in a sample are due to variations In variables that were not measured. We know of no reDable criterion for decisions of this kind in the social scientific or statistical fiterature. Work on Principle III yields one.

6.2 Tetrad Constraints from Acyclic Directed Graphs

The analysis of constraints implied by models goes back almost to very origins of covariance analysis. Besides vanishing partial correlations of various orders, another kind of constraint, vanishing tetrad differences, has been frequently considered in the psychometric and social science literature. A tetrad constraint is just a vanishing difference of products of correlations involving four distinct variables:

$$P_{ij}P_{kl} - P_{ik}P_{jl}$$

When does a directed graph (with or without unmeasured variables) imply a tetrad constraint? In Glymour, Scheines, Spirtes and Kelly an algorithm was given to determine the tetrad constraints implied by a graph, assuming the variables are linearly related. But that is a very stringent assumption. Spirtes has recently found a necessary and sufficient condition for a probability distribution that is representable by a directed acyclic graph using Pearl's formalism (Principle III)

Using Pearl's condition, Principle III, Spirtes has recently given a distribution-free necessary and sufficient condition for a graph to represent a distribution with a vanishing tetrad difference.

6.3 A sufficient condition for latent variables

A set of vanishing partial correlations can entail a vanishing tetrad difference. In causal models in which all of the variables are measured, all constraints implied by the graph of the model are axiomatized by the statistical dependencies and independencies represented by the graph, and all of the implied vanishing tetrad differences are implied by vanishing partial correlations. This observation led us [Glymour 87] to suggest a heuristic for introdudng latent variables: if a tetrad constraint holds that is not implied by vanishing first-order partial correlations among the measured variables, introduce latent variables. The theoretical grounds for a generalization of this heuristic can now be obtained rigorously as a corollary of Spines' characterization of the representation of vanishing tetrad differences:

Theorem III: Suppose probability distribution P Implies a vanishing tetrad difference

PijPkl " PikPjl

Then a directed acyclic graph G represents P In accordance with Principle III only If there exists a set S of vertices of G such that the variables I and j are Independent conditional on S, and so are the pairs of variables k, I and I, k and j, **I.**

When a vanishing tetrad constraint is judged to hold in the population, but there is no set of /Treasured variables that the variable pairs i, j, and k, I, and i, k, and j, I are conditionally independent on, it does not follow that the distribution on the measured variables is generated from a probability distribution including latent variables. What follows is that if the distribution on the measured variables is a restriction of any distribution represented by an acyclic directed graph in accordance with Principle III, then that directed graph must have latent variables.

7. Asymptotic Reliability

The asymptotic properties of procedures that infer causal structure from constraints exemplified in sample data depend on two things.

First, on the asymptotic properties of the procedures for determining that a set of constraints is satisfied;

Second, on the relation between directed graphs and sets of constraints satisfied in a population.

Asymptotically distribution free tests are available both for vanishing tetrad differences and vanishing partial correlations. Sample correlations may, however, be correlated as may sample tetrad differences, so that the properties of decisions made on a series of individual tests from samples may be different from the properties of dedsions made simultaneously for collections of constraints. We doubt that this is a serious problem in the limit, but we have no proof.

We have also proved a theorem that informally expressed, states that for any "natural" probability distribution on the linear coeffidents and variances of linear models based on a given causal

graph, the likelihood is zero that there will be a tetrad or other independence constraint that holds because of the coefficients and variances rather than because of the causal structure.

8. Conclusion

We believe the results described here, together with the TETRAD II results reported elsewhere, take a long step towards automated causal inference. A lot of work remains before a feasible robust general tool is available, however. The algorithm using Principle II needs to be implemented; the criterion for judging the presence of latent variables must be implemented; an automatic elaboration procedure using the strategy of TETRAD II needs to be implemented; more robust statistical tests of conditional independence need to be introduced; and the entire package needs to be tested extensively on simulated data. Combined with TETRAD II procedures for finding elaborations of an initial incomplete latent variable model, and with commercial packages for estimation and testing under a variety of distributional assumptions, major parts of the process of model construction can be automated, and automated with guarantees of reliability.

There is an important gap in the work we have described. Although we possess procedures for specifying causal models when statistical dependencies are not due to unmeasured common causes, and we possess means to determine when common causes are acting, and we possess means to finding omitted causal dependencies in a partially specified latent variable model, we do not possess a reliable informative procedure for constructing initial latent variable models from the data alone, nor do we possess a solution to the closely connected problem of feasibly generating all latent variable causal models statistically equivalent to a given causal model. These problems will require further fundamental research.

References

Duncan, O.D., *Introduction to Structural Equation Models,* Academic Press, New York, 1975

Glymour, C, Scheines, R.f Spirtes, P., and Kelly, K.f *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling,* Academic Press, San Diego, 1987.

Harary, F., and Palmer, E., *Graphical Enumeration,* Academic Press, New York, 1973.

James, L.R., Mulaik, S.A., and Brett, J., *Causal Analysis,* Sage Publications, 1982

Joreskog, K., and Sorbom, D., *LISREL VI: User's Guide,* Scientific Software, Mooresville, IN, 1984

Lee, S., Ph. D Thesis, Ohio State University, 1987

Loehiin, J., *Latent Variable Models,* Lawrence Ehrlbaum Associates, Hillsdale, NJ, 1987.

Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann Publishers Inc., San Mateo, CA, 1988

Rodgers, R.C., and Maranto, C.L, Causal Models of Publishing Productivity in Psychology, *Journal of Applied Psychology,* Vol. 74, No. 4, pp.636-649., 1989.

Skyrms, B. *Causal Necessity,* Yale Universiyty Press, New Haven, 1980.

Spirtes, P., Scheines, R., and Glymour, C, Simulation Studies of the Reliability of Computer Aided Model Specification Using the TETRAD II, EQS, and LISREL Programs, forthcoming in *Sociological Methods and Research,* Sage, 1989.

Spirtes, P., A Necessary and Sufficient Condition for Conditional Independencies to Imply a Vanishing Tetrad Difference, Technical Report Number CMU-LCL-89-3, Dept. of Philosophy, Carnegie Mellon University, Piitsburgh, PA, 15213.

Suppes. P. *A Probabilistic Theory of Causality,* North-Holland. Amsterdam. 1970.