Reply to Freedman

by Peter Spirtes and Richard Scheines

In Causation, Prediction, and Search (Spirtes, Glymour, and Scheines 1993), we undertook a three part project. (Henceforth we will refer to Causation, Prediction, and Search as CPS.) First, we characterized when causal models are indistinguishable by population conditional independence relations under several different assumptions relating causality to probability. Second, we proposed a number of algorithms that take sample data and optional background knowledge as input, and output a *class* of causal models compatible with the data and the background knowledge; the algorithms (with the exception of the heuristic algorithm described in Chapter 11) were accompanied by proofs of their correctness given assumptions that were clearly stated in CPS, and that we will restate below. Finally, we offered a theory of how to predict the effects of interventions in causal structures, given only partial knowledge of causal structure. Freedman's objections are all directed against the causal inference algorithms we proposed. We do not have room here to discuss all of his criticisms, but we have answered his major points. With regard to the points we do not have room to discuss, the reader should be warned that Freedman is an unreliable interpreter of what we have written. For convenience, we have divided Freedman's objections into the following categories.

1.) Freedman questions some of the assumptions on which our correctness theorems are based. Some of his criticisms are based on covariance matrices that he constructed. None of the examples he constructed in sections 11.2, 11.3, or 12.3 are counterexamples to any theorem that we stated, nor are they even germane to the question of how probable are the assumptions we make. His examples only illustrate points discussed in detail in our book (particularly in the chapter on indistinguishability), in which we give similar examples.

2.) The most serious charge that Freedman makes is that the algorithms do not compute what we say they do. According to Freedman:

> SGS seem to claim that their algorithms will find all the path diagrams compatible with a given covariance matrix. However, the theorems suggest that the algorithms will at best find one such graph. SGS also seem to claim that their algorithms are consistent. However, without an identifiability theory for linear models, they cannot be talking about consistency. (Freedman, 49)

Freedman's claim that our algorithms do not find all of the path diagrams "compatible" (in a sense explained in detail below) with a given covariance matrix is false, and based on a fundamental misunderstanding of the output of the algorithms and the theorems of correctness.

3.) Finally, Freedman claims our empirical cases our unconvincing. He raises many substantive objections to a number of the models of real data that were presented in our book. What we said about these examples was:

> We illustrate the algorithms for simulated and real data sets. With simulated data the examples illustrate the properties of the algorithms on samples of realistic sizes. In the empirical cases we often do not know whether an algorithm produces the truth. But it is at the least very interesting that in cases in which investigators have given some care to the treatment and explanation of their data, the algorithm reproduces or nearly reproduces the published accounts of causal relations. It is also interesting that in cases without these virtues the algorithm suggests quite different explanations from those advocated in published reports. (Spirtes, Glymour, and Scheines 1993, 132-133.)

Freedman's substantive criticisms of the models produced do not in any way contradict this claim.

Freedman's criticisms are marred by some fundamental misinterpretations of what we say. On more than one occasion he criticizes what we "seem" to say, even though we have never held or stated the position he attributes to us. Of the few direct quotations from us that he gives, two are taken out of context (Freedman, 24 and 25). In one of these two quotations, Freedman replaces key qualifiers which completely change the meaning of the quotation with "****", and remarks that "Readers may consult the original for context, to see whether the omitted material affects the meaning." (Freedman, 24) Later we will produce the full quote so that readers will be able to see how the omitted material and the context affects the meaning.

1. The PC Algorithm

A number of Freedman's objections are based on misunderstandings of the algorithms that we proposed and their accompanying correctness theorems. Hence, it will be useful to give brief descriptions of the algorithms and the correctness theorems here. Since most of Freedman's objections are directed against the PC algorithm, that is the algorithm we will discuss in the most detail.

A directed acyclic graph (DAG) can be used to represent two distinct types of objects. First, it can be used to represent a set of probability distributions, all of which share certain conditional independence relations in common. Second, it can be used to represent causal relationships between variables; given a causally sufficient set of variables $\mathbf{V}$ (i.e. every common cause of variables in $\mathbf{V}$ is also in $\mathbf{V}$), there is an edge from X to Y in G if and only if X is a direct cause of Y relative to $\mathbf{V}$. We call DAGs interpreted in this way causal DAGs. These two distinct uses of DAGs can be linked by the following assumptions:

> **Causal Markov Condition:** Let G be a causal DAG with vertex set $\mathbf{V}$ and P be a probability distribution over the vertices in $\mathbf{V}$ generated by the causal structure

represented by G. G and P satisfy the Causal Markov Condition if and only if for every W in **V**, W is independent of $\mathbf{V}\backslash(\mathbf{Descendants}(W) \cup \mathbf{Parents}(W))$ given $\mathbf{Parents}(W)$.

Faithfulness Condition: Let G be a causal DAG and P a probability distribution generated by G. $<G, P>$ satisfies the Faithfulness Condition if and only if every conditional independence relation true in P is entailed by the Causal Markov Condition applied to G.

Their meaning and justification are discussed more fully in Glymour (this volume) and in CPS. However, if one makes these assumptions, then the probability distribution of data generated by a causal process represented by causal DAG G is a member of the set of probability distributions represented by G.

Sometimes different DAGs represent the same set of probability distributions. In that case we say that the two DAGs are faithfully indistinguishable. (In CPS we define two DAGs as faithfully indistinguishable when they entail the same set of conditional independence relations. But if they entail the same set of conditional independence relations they represent the same set of probability distributions.) For example, the following two DAGs are faithfully indistinguishable.

Figure 1

For each DAG G, there is a set of DAGs that are faithfully indistinguishable from it; we call this the faithful indistinguishabilty class of G. The faithful indistinguishability class of (i) in Figure 1 consists of the DAGs (i) and (ii). This faithful indistinguishability class can be more compactly represented by an object that we call (following Verma and Pearl 1992) a **pattern**. Hence a pattern represents a set of DAGs (which in certain special cases may contain only one member). A pattern is a type of graph that can contain both undirected edges and directed edges. The pattern representing the faithful indistinguishability class of (i) in Figure 1 is shown in Figure 2.

Figure 2

We will not explain here the rules for determining whether a given DAG G is in the set of DAGs represented by a pattern. See CPS, and Glymour (this volume) for more details. However, notice that when an oriented edge (e.g X -> Y) in figure 2 appears in a pattern, then the edge X -> Y occurs in *every* graph in the set represented by the pattern. In contrast, when an unoriented edge (e.g. Z - W) appears in a pattern, then the two variables it connects are adjacent in every DAG in the set represented by the pattern, but they may have different orientations (e.g. Z -> W in (i) and W -> Z in (ii)).

Since all of the members of a faithful indistinguishability class represent the same set of probability distributions, it is not possible to determine which member of the class generated a given distribution in the set. For that reason, the output of the PC algorithm is not a single DAG, but a faithful indistinguishability class represented by a pattern, although in some special cases, the set of DAGs represented by a pattern contains a single DAG.

The PC algorithm takes as input sample data, and optional background knowledge. The algorithm relies on tests of conditional independence, so in practice a user also has to assume something about the class of distributions that the population distribution lies in. In implementations of the algorithm, we have allowed the user to assume that the population is either linear normal or discrete. Freedman dislikes the assumption of linearity in many cases, but this assumption is not essential, as long as some reasonable distributional assumption can be made which enable the algorithm to make judgments about conditional independence. Of course, the distribution may either be completely unknown, or so complicated as to make judgments about conditional independence impossible in practice.

The correctness theorem for the PC algorithm states that assuming (i) the Causal Markov Condition, (ii) the Faithfulness Condition, and (iii) causal sufficiency (i.e. every common cause of a pair of measured variables is itself measured), then from conditional independence relations true in the population, the PC algorithm outputs a pattern that represents the faithful indistinguishability class of the true causal DAG. For example, given a distribution generated by the graph (i) in Figure 1, the output of the algorithm is the pattern shown in Figure 2.

2. Assumptions

According to Freedman:

> Their algorithms [those found in CPS] have some technical interest, but will make causal inferences only when causation is assumed in the first place. To be more explicit: If we assume that the arrows in a path diagram represent causation rather than association, and we also assume that the path diagram can be estimated from data, then indeed SGS can infer causation from association. ... Even so, causality is assumed into the picture at the beginning, not proved in at the end. (Freedman, 65)

It is not clear what this passage means. The output of the PC algorithm is a pattern, typically not a path diagram (or DAG). We do simply assume that the data is generated by *some* causal process. We do not simply assume that the output of the algorithm (given correct judgments about population conditional independence relations) represents a set of DAGS, one of which is a correct description of that causal process. We *derive* that conclusion from the Causal Markov and Faithfulness conditions, and causal sufficiency. (If the assumption of causal sufficiency seems unreasonably in a particular case, the FCI algorithm we proposed can be used in place of the PC algorithm.) Moreover we believe that these axioms are often widely, but not explicitly, already assumed in a variety of contexts by practicing statisticians.

Freedman complains that "causality is assumed into the picture at the beginning." It is a banality that drawing causal inferences from statistical data requires making assumptions about the relationship between statistical data and causal structure. Such assumptions are even made by experimentalists (who assume an instance of the Causal Markov Condition.) We fail to see the point in replacing a discussion of two precisely stated axioms, the Causal Markov and Causal Faithfulness Conditions, with complaints about something so vague as "causality is assumed into the picture at the beginning."

Instances of the Causal Markov Condition are assumed in a wide variety of contexts in statistics. Since Freedman does not raise objections to the Causal Markov Condition, we will not discuss it further here.

The Faithfulness Condition is essentially an assumption that conditional independence relations are generated because of the structure of the DAG (i.e. for all parameterizations of the DAG), rather than because of the parameters (i.e. for some, but not all, parameterizations of the DAG.) In CPS (Chapter 3) we discuss a number of ways in which the Faithfulness Condition can fail. One way it can fail is if there are deterministic relationships among measured variables. The path models in figure 3 illustrate another way in which the Faithfulness Condition can fail (where the error terms have been omitted from the DAGs.)

Figure 3

$$X = \varepsilon_X$$

$$Y = \mathrm{a} \times X + \varepsilon_Y$$

$$Z = \mathrm{b} \times Y + \mathrm{c} \times X + \varepsilon_Z$$

Suppose that the data were generated by a linear causal process described by DAG (i), and the error terms are independent normally distributed variables. This does not entail any conditional independence relations among the variables for all values of the linear coefficients. However, if $\mathrm{c} = -\,\mathrm{a} \times \mathrm{b}$, then X and Y are independent. If the PC algorithm were applied to this data, it would incorrectly produce a pattern whose only member is model (ii). The distribution is not faithful to DAG (i) which generated it, but is faithful to DAG (ii) which did not. This is not a violation of the theorem we proved about the correctness of the PC algorithm, because the theorem assumes that the distribution is faithful to the DAG of the causal structure that generated the data (DAG (i)) which is not true in this case. The examples that Freedman constructs in his section 11.3 adds nothing to our discussion in CPS of indistinguishability, and the examples in 11.2 and 12.3 add nothing to our discussion in CPS of faithfulness.

However, the mere existence of cases where the Faithfulness Condition fails is of no interest in itself. To draw an analogy, estimation algorithms can fail in the large sample limit on sets of measure 0; this is not an objection to the estimation methods. Similarly, the relevant question here is not the existence of data sets on which the Faithfulness Condition will fail, but how likely it is that it will fail. The examples that Freedman gives in sections 11.2, 12.3 don't even address this issue.

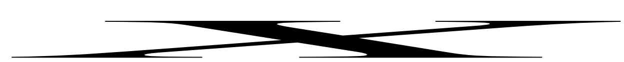
In the linear case, for a given directed DAG G, if we parameterize the space of models by the linear coefficients and the variances of the exogenous variables, the probability of the Faithfulness Condition failing in the population is of Lebesgue measure 0. This has the consequence that any Bayesian who, conditional on belief in a causal structure represented by G, has a distribution over the linear parameters that is absolutely continuous with Lebesgue measure, assigns measure 0 to a violation of the Faithfulness Condition. Of course we cannot prescribe what priors a Bayesian ought to have, but we think that the kind of prior that we have described is implicitly held by many people in many cases.

James Robins (personal communication) has raised questions of a more interesting type. We will consider here one simple variation of the kind of question he asks. Suppose that from a sample the population is judged to be faithful to G_2 but not G_1. One would expect that the probability of G_2 given the data would be boosted relative to the probability of G_1 given the data, because while *all* parameterizations of G_2 are compatible with the conditional independence relations judged to hold in the population, only a relatively small set of parameterizations of G_1 are compatible. But how large is this boost in probability, and might it not be overwhelmed by priors favoring G_1 over G_2? To test this, we did the following Bayesian analysis. We generated 2000 sample data points from a random parameterization of DAG (ii) in Figure 3, where each of the variables is ternary. The PC algorithm correctly found that DAG (ii) generated the data. We assigned equal prior probabilities to DAG (i) and DAG (ii), and we assigned a Dirichlet distributions to the parameters of DAG (i) and DAG (ii). (The distribution we used assigns the same probability to any two faithfully indistinguishable DAGs and is described in Heckerman, Geiger, and Chickering 1994) We then calculated the ratio of the posteriors of DAG (i) to DAG (ii) given the data for 10 different data sets generated from random parameterizations of (ii) where each variable could take on three different values, and 10 different data sets generated from random parameterizations of (ii) where each variable could take on four different values. (We did the calculations using the K2 program lent to us by Cooper.) The likelihood ratio of DAG (ii) to DAG (i) given the data for cases where each variable took on three different values averaged $10^{6.21}$, and for the cases where each variable took on four different values it averaged $10^{13.7}$. For the case where the true model is (ii), this gives some indication of how heavily the priors would have to favor G_2 before we would be justified in favoring the alternative to the output of the PC algorithm. More Bayesian analyses of this kind, using different graphs, latent variables, and different priors over the parameters would be useful, and we plan to perform them.

3. The Correctness Results

The correctness theorem for the PC algorithm states that assuming the data is generated by some causal DAG, the Causal Markov and Faithfulness Conditions, and causal sufficiency (i.e. every common cause of a pair of measured variables is itself measured), then the PC algorithm outputs a pattern that represents the faithful indistinguishability class of the true causal DAG as long as its judgements about conditional independence relations in the population are correct. Although the theorem was stated under the assumption that population conditional independence facts were known it is easy to show that the output of the PC algorithm is correct with probability 1 in the large sample limit, if instead of being given population conditional independence relations, one is given a class of distributions that the distribution over the measured variables falls in, and a test of conditional independence in that class of distributions which is correct with probability 1 in the large sample limit.

According to Freedman:

> SGS seem to claim that their algorithms will find all of the path diagrams compatible with a given covariance matrix. However, the theorems suggest that the algorithms will at best find one such DAG. (Freedman, 49)...
>
> In statistical terminology, models are "identifiable" when they make different predictions about observables. For example, suppose you have two models for your data. If, for all data sets
>
> $$P(\text{data} \mid \text{model } 1) = P(\text{data} \mid \text{model } 2),$$
>
> there is an obvious problem-- the data cannot distinguish between the models. If a path model is complete, or the faithfulness assumption is not imposed, the graph underlying a covariance matrix is not identifiable; ... However, even if we assume that a covariance matrix is faithful to a graph that is not complete, there may be several such graphs (SGS p. 89). For example, the following three graphs can generate the same covariance matrix:
>
> 
>
> Thus, SGS do not seem to have succeeded in defining a class of graphs and covariance matrices for which identifiability holds... (Freedman, 49)

This passage reveals several confusions that Freedman has about the output of the algorithm. First, we never claimed to be able reliably determine a unique DAG from a covariance matrix. That is why if the data is generated by any of the three causal DAGs Freedman refers to, the output of the algorithm is a pattern that represents *all three* DAGs. Moreover, each DAG in the set represented by the pattern is not a statistical model in the usual sense. That is, a DAG is not associated with one probability distribution, it is instead associated with a set of probability distributions (which under the assumption of faithfulness all share the same set of conditional independence relations.) It is only after a DAG is parameterized that it is associated with a single probability distribution. So the concept of identifiability as defined by Freedman does not even apply to the graphs that are represented by the output of the PC algorithm; P(data | graph) is not well defined. We were of course well aware of the fact (and repeatedly emphasized) that even under the assumption of faithfulness, that there exist *parameterizations* of the three graphs in Freedman's example that represent the same covariance matrix. It is a simple consequence of a number of theorems described in Chapter 4 of CPS (from Verma and Pearl 1990, and Frydenberg 1990.) It is for precisely this reason that the output of the PC algorithm in this case is not a single DAG, but a pattern which represents all 3 of the DAGs. And as previously explained, we *have* identified sets of models which cannot be distinguished from each other given just a probability distribution, assuming Causal Markov, Faithfulness, and causal sufficiency; this is just the faithful indistinguishability class of a DAG.

In section 11.3 Freedman constructs examples of models that cannot be distinguished by their correlation matrices. All of the examples of indistinguishability that he presents in this section are simple consequences of the indistinguishability theorems we state in chapter 4 of CPS. Moreover, if data from any of the indistinguishable models were given to the FCI algorithm, its output would be a set of graphs that contained all of the indistinguishable models.

The relevant questions about the output of the PC algorithm are:

1. Is the set of DAGs represented by the output of the PC algorithm too small (in the sense that it might leave out the true DAG)? The answer is no, if one assumes Faithfulness, Markov, causal sufficiency, and correct statistical decisions about conditional independence.

2. Is the set of DAGs represented by the output of the PC algorithm too large (in the sense that, under the assumptions made, some DAG represented by the output pattern does not have a parameterization that faithfully generates the true distribution, and hence could not be correct?) Again the answer is no. (In CPS we showed that each of the DAGs represented by the output pattern entailed the conditional independence relations that hold in the true distribution. Under the assumption of causal sufficiency, and the existence of the right conditional density functions, this entails the stronger result that each of the graphs has a parameterization that faithfully generates the true distribution.)

There are parallel correctness results for the other algorithms in CPS (except for the algorithm described in Chapter 11, which is heuristic.)

In an artfully worded passage, Freedman hints that despite our claims, we ultimately have disavowed the possibility of discovering causal relations from conditional probabilities. He says:

> SGS discuss the role of assumptions in their theory several times (pp. 53-69, pp. 75-81, pp. 324-5, p. 351). However, the clearest statement can be found when SGS are trying to discredit the evidence that smoking causes lung cancer:
>
> > effects **** cannot be predicted from **** sample conditional probabilities. [p. 302]
>
> Readers may consult the original for context, to see whether the omitted material affects the meaning. The advantage of the quote is clarity. If the statement is

> generally applicable, then SGS-- like Yule and Pearl before them-- have disavowed the ability to infer causation from association. (Freedman, 24)

Here is the quote with the omitted words restored, and some more of the beginning of the sentence:

> ...leading epidemiologists, such as Lillienfeld, seem simply not to have understood that *if the relation between smoking and cancer is confounded by one or more common causes*, the effects of abolishing smoking cannot be predicted from the "risk ratios," i.e., from sample conditional probabilities. (Spirtes, Glymour, and Scheines 1993, 302, italics added)

The conversational implication of the words removed is quite clear; if the relation between smoking and lung cancer is *not* confounded by one or more common causes, the effects of abolishing smoking *can* be predicted from sample conditional probabilities. Moreover, this latter claim is a simple application of a theory developed at length in Chapter 7 of CPS; the goal of that theory was to specify conditions under which it is possible to calculate the effects of interventions from conditional probabilities, and the conditions under which it is not. In the context of the theory developed in Chapter 7, and with the omitted words are restored, it is clear that the quotation is an *application* of our theory, not a *denial* of it. (Incidentally, the Freedman quotation reveals another fundamental confusion of Freedman's: the quotation concerns the possibility of making quantitative predictions of the effects of causal intervention from partial knowledge about causal structure and sample probability distributions, not the possibility of inferring features of causal structure from sample probability distributions.)

4. Examples

At the beginning of the discussion of empirical examples we said:

> We illustrate the algorithms for simulated and real data sets. With simulated data the examples illustrate the properties of the algorithms on samples of realistic sizes. In the empirical cases we often do not know whether an algorithm produces the truth. But it is at the least very interesting that in cases in which investigators have given some care to the treatment and explanation of their data, the algorithm reproduces or nearly reproduces the published accounts of causal relations. It is also interesting that in cases without these virtues the algorithm suggests quite different explanations from those advocated in published reports. (Spirtes, Glymour, and Scheines 1993, 132-133.)

We do not think that Freedman's substantive criticisms in any way contradict this claim. We do not have room to discuss all of the criticisms, but we will discuss some of the major ones here.

4.1 The Timberlake-Williams example

The point of our discussion of the Timberlake-Williams case was that the FCI algorithm produced an alternative to the Timberlake-Williams model that was consistent with background knowledge, was testable and fit the data. We think that this sheds considerable doubt upon the Timberlake-Williams model. None of Freedman's remarks are relevant to our criticism of the Timberlake-Williams model. Freedman does not think any of the models discussed (including the one constructed from the output of the FCI algorithm) have good evidence for them, because some of the assumptions are dubious, and the sample size is small. We agree, but we never made any claim to the contrary. We even pointed out that we could not reproduce the correlation matrix reported by Timberlake and Williams from the sources they cite. Our own simulation studies indicated that the PC algorithm (which is considerably simpler than the FCI algorithm) was not very reliable at samples sizes considerably larger than that in the Timberlake-Williams studies.

4.2 The AFQT example

We were given a variety of test scores (which we will call the subtest scores) and an AFQT score. The AFQT score is calculated from some of the subtest scores we were given, but not others. The point of this example was to show that *given the same background knowledge*, regression techniques could not find the components of the AFQT score, but the PC and FCI algorithms can. Freedman points out that given *more* background knowledge than was given to the PC and FCI algorithms, regression techniques would be able to pick out the components of the AFQT score. (In particular, his regression analysis needs two further assumptions: first that an increase in a component score increases the AFQT score, and second that a component score is not highly negatively correlated with another component score because of a common cause.) Both of these assumptions are reasonable in this case. But it does not take away from the fact that the PC and FCI algorithms derive the same result with less information, and in cases where these background assumptions are not so reasonable, the regression technique could not be employed. Freedman also points out although the PC algorithm gets the right relationship among the subtest scores and the AFQT score, it gets the wrong relationship among the subtest scores. This was evident because the PC algorithm output a pattern that represented graphs that had a cycle, which does not make substantive sense in this case. This is true. However, this application of the algorithm assumed linearity. The relationship of the component scores to the AFQT score (as calculated in 1987) is *known* to be linear. There is no reason at all to believe that the relationship among the subtest scores is linear. So, where the distributional assumptions are known to be correct, the algorithm gets the right answer, and where there was no reason to believe that the assumption is correct, the algorithm gets the wrong answer.

4.3 Smoking and Lung Cancer:

According to Freedman:

> ... does smoking cause lung cancer, heart disease, and many other illnesses?... When they actually get down to arguing their case, they [SGS] use a rather old-fashioned method-- a literature review with arguments in ordinary English (pp. 291-302). Causal models and search algorithms have disappeared. (Freedman, 26).

> SGS elected not to use their analytical machinery on the smoking data-- a remarkable omission. (Freedman, 27).

The output of a search algorithm is one kind of evidence for a causal claim, but there are others. We discussed some of them in the case of smoking and lung cancer. The reason that we discussed the smoking and lung cancer case in ordinary English was because we felt a historical account of that case concretely illustrated how the issues concerning causal inference from observational data, and prediction of the effects of intervention had been dealt with by the statistical community. Freedman disagrees with some of our discussions of the evidence. The relevance of any of this to the question of the reliability of the algorithms we propose escapes us.

We did not apply the algorithms to smoking and lung cancer data because we happened not to possess any such data. Without the data, we do not know which variables were measured, how they are distributed, etc. Whether the algorithms are applicable and would produce an interesting result depends upon the answers to these questions. But there is nothing in principle that prevents the application of the algorithms to such data.

4.4 The Rindfuss Example

Freedman points out there were some errors in the covariance matrix we used. He does not point out that those errors make no difference in the output of the algorithm.

5. Conclusion

Much of applied statistics is devoted to drawing causal conclusion from statistical data. But when examining causal inferences, statisticians all too often simply trot out a few examples of circumstances under which particular causal inferences cannot be made. What is needed is a rigorous examination of what causal relations can be inferred given varying kinds of background knowledge. What could we infer about causal structure if we knew the time order of the variables? What if we knew no common causes were acting? What if we have measured a large number of variables? In addition, there should be a theory about when is it possible to determine the effects of an intervention upon that causal structure, given the inferences that can be made about causal structure. CPS attempts to systematically examine these kinds of questions. Under a variety of assumptions there are positive results that state when two different causal structures can be distinguished given the data, and negative results that state when they cannot. There is a systematic examination of the relationship between partial knowledge about causal structures and what can be predicted about the effects of intervention. It is not complete, and there may be many other kinds of assumptions that should be investigated, but we think that it is an example of the kind of research into causal inference that should be an important part of statistics.

The space of possible causal models, even when many have been eliminated by background knowledge, is far too large to search without automated help. One part of our project was the proposal of some algorithms that use statistical data and background knowledge supplied by a user to suggest causal models compatible with the background knowledge and the data. As we have consistently emphasized (despite Freedman's claims to the contrary) it is an empirical question whether the algorithms are reliable when applied to real data. Even if the algorithms are useful in practice, informative and reliable causal inference from correlational data will not be easy. The right variables must be measured, they have to be distributed in such a way that reasonable tests of conditional independence can be made, and the causal relationships among the variables cannot be too complex. All of these caveats are clear in our book and emphasized in Glymour (this volume).

There are three kinds of evidence concerning the reliability of the algorithms. First they are asymptotically correct given the assumptions we make. Second, simulation tests indicate that when the modeling assumptions are satisfied, and the sample sizes are reasonably large, they are reliable. Third, there is some evidence from application to real data sets.

The evidence from application to real data sets is the most fragmentary. We generally did not have access to raw data, and in most cases we do not know whether the output, however reasonable it may seem, is correct. In the AFQT case, among the relationships known to be linear, it output an answer known to be correct. In the case of Spartina biomass (Spirtes, Glymour and Scheines 1993, 244-248), the algorithm again predicted from observational data results confirmed by experiment. In a number of other examples in CPS that were not discussed by Freedman, the answers appeared reasonable, but it is not known whether they are correct. More extensive applications will provide more evidence about the reliability of the algorithms to real data sets. We are currently carrying out some applications of the algorithms to medical databases.

References

Frydenberg, M. (1990). "The Chain Graph Markov Property." *Scand. J. Statist,* **17**, 333-353.

Heckerman, D., Geiger, D., and Chickering, D. (1994) Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, Technical Report MSR-TR-94-09, Microsoft Research, Advanced Technology Division, Redmond WA.

Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation, prediction, and search. Lecture Notes in Statistics, 81, Spring-Verlag, New York.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. Proc. Sixth Conference on Uncertainty in AI. Association for Uncertainty in AI, Inc., Mountain View, CA, 220-227.

$$X \longrightarrow Y \longleftarrow Z \longrightarrow W$$

(i)

$$X \longrightarrow Y \longleftarrow Z \longleftarrow W$$

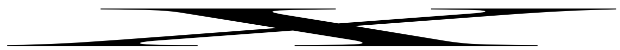
(ii)

(i) (ii)

$$X \longleftarrow Y \longrightarrow Z$$

$$\begin{tikzcd} X \arrow[r, dash] & Y \arrow[r, dash] & Z \end{tikzcd}$$

(i)



(ii) (iii) (iv)