in contemporary statistics is not the only way to achieve simplicity. The main lesson that I took away from Wermuth's doctoral research (cf. Dempster, Schatzoff and Wermuth, 1977) is that smooth systems of declining parameter values are usually a more efficient way to simplify statistical complexity than sharp cutoffs that set most parameter values to zero. Computational strategies of choice then become radically different. Classical estimation techniques that are adequate with relatively few parameters must be replaced with Bayesian or similar methods that reflect prior assessments of patterns of smooth decline. Donoho et al. (1992) illustrate a notable non-Bayesian approach. My own preference is for Bayesian models with many more hidden variables and many more dependence parameters than SDLC allow, to have a reasonable possibility of capturing actual mechanisms. I believe that rapidly developing computing power and algorithms that sample posteriors should be used to implement and test more complex Bayesian models.

Beyond the elicitation of priors and beyond the problem of simplifying the complex structures of highly multivariate and selectively filtered populations encountered in real practice, there remains a gray area that SDLC address briefly in two sentences as situations where "the number of assessments made is insufficient to specify a joint distribution uniquely." The use of maximum entropy or other arbitrary prior generation principles typically leads to exactly the unrealistic procedures that the smoothing of large parameter sets is designed to avoid. SDLC fail to mention the belief function approach (Shafer, 1976) that Dempster and Kong (1988) show fits naturally into network modelling built on decompositions of evidence into independent sources similar in spirit to the "graphical modelling" approach of SDLC. It is my view as a coinventor of the BEL theory that it is a near cousin of the Bayesian strategy that descends directly from classical subjective probability and is not a foreign interloper from distant tribes of semicoherent formal systems. Unlike the naive upper and lower probability models that have been studied by Good, Walley and others, the BEL system constructs models from judgmentally independent assessments on knowledge spaces and combines the components by a simple precise rule that reduces to the Bayesian rule for combining likelihood and prior in the special Bayesian case. The chief hindrance to developing and testing BEL models for probabilistic expert systems has been computational difficulties. Shafer, Kong and others showed in the mid-1980s how to decompose BEL computations coincidentally with the parallel demonstrations of Lauritzen and Spiegelhalter (1988) that SDLC feature. But these clever algorithms only stave off computational complexity temporarily. The future of both Bayesian and BEL approaches depends on the revolution that has been gathering speed for the past five years on Monte Carlo posterior sampling.

ACKNOWLEDGMENTS

I thank Emery Brown for helpful discussions on both medical and statistical issues. My work is partially supported by ARO Grant DAAL03-91-0089 and NSF Grant DMS-90-03216.

Comment: Conditional Independence and Causal Inference

Clark Glymour and Peter Spirtes

Fourteen years ago, in an essay on conditional independence as a unifying theme in statistics, Philip Dawid wrote that "Causal inference is one of the most important, most subtle, and most neglected of all the problems of Statistics" (Dawid, 1979a). Only shortly later, several statisticians (Wermuth and Lauritzen, 1983; Kiiveri and Speed, 1982) introduced frameworks that connect conditional independence, directed acyclic graphs (hereafter DAGs) and causal hypotheses. In these models the vertices of a DAG G represent variables, and a directed edge $X \rightarrow Y$ expresses the proposition that some change in variable X will produce a change in Y even if all other variables represented in G are prevented from changing. The power and generality of DAG models derive from their dual role in representing both causal or structural claims and

Clark Glymour is Alumni Professor of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, and Adjunct Professor of History and Philosophy of Science, University of Pittsburgh. Peter Spirtes is Associate Professor of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

also patterns of independence and conditional independence constraints on distributions. The paper by Spiegelhalter, Dawid, Lauritzen and Cowell (SDLC) provides a valuable review of the current state of the art in using and constructing statistical and causal hypotheses represented by DAGs. The paper by Cox and Wermuth (CW) lays out interesting problems concerning how to generalize DAG models. Our remarks concern four issues raised – explicitly or implicitly – by these papers:

1. Do other graphical objects with a plausible causal or structural interpretation represent sets of conditional independence relations that cannot be represented by DAGs? We will briefly describe generalizations of DAG models representing marginals of distributions with latent variables and generalizations representing feedback; graphical chain models do not represent such processes.
2. Are classifications or diagnoses using Bayesian networks or DAG models more reliable than those made by other existing classification techniques? We believe the question is unsettled.
3. Besides classification, what other uses do DAG models have? We think the essential use of such models is in predicting the effects of interventions – experiments, policies, etc. – that change the joint distribution of variables in a population, and this use connects these models with analyses by Rubin (1974, 1977) and others of the invariance of conditional probabilities under interventions and with a wealth of issues in experimental design.
4. What is the state of the art of automatic techniques for constructing DAG models? We will briefly note properties of several procedures that appear to be more generally applicable than the automated search illustrated in the SDLC review.

1. DIRECTED ACYCLIC GRAPHS AND GENERALIZATIONS

After introducing a variety of graphical structures to represent patterns of conditional independence relations not represented by any DAG – "nondecomposable" sets of conditional independence relations – and illustrating these patterns in empirical examples, CW say: "Our examples illustrate that such nondecomposable structures arise in different contexts. There is need to identify them and to find explanations of how they could have been generated." This question, as we understand it, asks what sorts of causal processes might lead to nondecomposable patterns of conditional independence relations; that the issue is posed near the end of their paper suggests that when nondecomposable patterns are found, the various graphical representations CW consider have no clear interpretation as causal hypotheses. To address their question, we first briefly consider the connection between causal structure and conditional independence in DAG models, then in graphical chain models and finally in alternative generalizations of DAG models.

In various frameworks, each DAG can be paired with any member of families of probability distributions over variables represented by vertices in the graph. The frameworks differ in their selection of restrictions on graph/distribution pairs, $\langle G, P \rangle$. Common restrictions include: (1) the Markov condition (Kiiveri and Speed, 1982): for admissible $\langle G, P \rangle$ X is independent of its nondescendants in G given its parents in G; (2) the "recursive diagram" or "directed independence graph" condition (Wermuth and Lauritzen, 1983): for admissible $\langle G, P \rangle$ and a given ordering of variables, $X \rightarrow Y$ is in G if and only if Y is after X in the ordering, and Y is dependent on X conditional on the set U of all vertices (excluding X) that precede Y in the ordering; (3) the Minimality condition (Pearl, 1988): for $\langle G, P \rangle$ satisfying the Markov condition, if H is a proper subgraph of G then $\langle H, P \rangle$ does not satisfy the Markov condition; (4) positivity of distributions; (5) the DAG isomorph or Faithfulness condition (Pearl, 1988): for admissable $\langle G, P \rangle$, vertices X, Y are independent conditional on set U of vertices only if the Markov condition applied to G entails that conditional independence. The restrictions on graph/distribution pairs are related. Directed independence graphs + positivity is equivalent to Markov + Minimality + positivity. Markov + Faithfulness + positivity entails the other conditions but is strictly stronger than Markov + the other conditions.

The Markov condition, the directed independence graph condition and the Minimality condition are directly motivated by intuitions about causality reflected in statistical practice throughout the century and in philosophy of science for almost half a century. [A few examples: a special case of the Markov condition is essential to Fisher's (1951) arguments in *The Design of Experiments* and throughout subsequent work on experimental design; the Markov condition is the guiding idea of latent variable models, as Bartholomew's (1987) recent review notes (without mentioning directed graphs explicitly); the Markov and Faithfulness conditions are tacitly assumed in the arguments about model selection developed by Simon (1954) and by Blalock (1961) early in the 1960s. In philosophy of science, aspects of the directed independence graph condition, for example, were given in a condition for "probabilistic causality" proposed by Suppes (1970) and aspects of the Markov condition were given by Reichenbach (1956).]

The Faithfulness condition can be viewed as requiring stability of conditional independence over small variations in parameters in models; in other terms,

conditional independence facts are to be explained by structure alone. Under a natural parameterization of linear normal models satisfying the Markov condition for a DAG, G, the set of unfaithful distributions has zero Lebesgue measure (Spirtes, Glymour and Scheines, 1993). The Markov and Faithfulness conditions are realized – sometimes without explicit graphical representations – in a wide array of models with causal interpretations in the social sciences, epidemiology and elsewhere and in the design of experiments and derivation of null hypotheses.

For every distribution P over a set of variables V and every ordering of the variables there exists a DAG compatible with the ordering such that P satisfies the Markov and Minimality conditions and a DAG compatible with the ordering such that P satisfies the directed independence graph condition. In contrast, there are many distributions that satisfy both the Markov and Faithfulness conditions for no DAG whatsoever; even if for some orderings of the variables there is a DAG for which P satisfies the Markov and Faithfulness conditions, there may not be such a DAG for every ordering of the variables. Unlike the other combinations of assumptions, the Markov and Faithfulness conditions jointly enable independencies to give some information about the directions of edges. The distributions that CW call "nondecomposable" do not satisfy the Markov and Faithfulness conditions for any DAG. Their introduction of other graphical representations for nondecomposable distributions therefore suggests that CW are implicitly imposing the Faithfulness condition on the set of distributions represented by a DAG.

1.1 Graphical Chain Models

CW describe a number of different kinds of "block" graphs, some of which represent sets of conditional independence relations that cannot be represented by DAGs unless the Faithfulness condition is violated. Their structures include graphical chain models in the sense of Lauritzen and Wermuth (1989) (hereafter, LW), structures also discussed by SDLC. These objects contain directed edges, undirected edges and variables grouped into blocks. The blocks of variables are linearly ordered; a directed edge $X \rightarrow Y$ occurs only if X is in a block previous to Y; undirected edges can only join variables in the same block. An edge $A \rightarrow B$ or $A - B$ occurs if and only if A and B are dependent conditional on the set of all variables occurring in the same block as B or in previous blocks.

The terminology of "explanatory" and "response" variables, and other remarks in the review papers, strongly suggest that directed edges in graphical chain models are given a causal interpretation, but the causal or structural significance of blocks and undirected edges is problematic. Wermuth and Lauritzen (1990) say little more than that variables joined by undirected edges in the same block are "on an equal footing." SDLC suggest undirected edges $X - Y$ represent reciprocal causation; in some units of the population X influences Y and in other units Y influences X. Under this interpretation, the chain graph represents a mixture of two subpopulations, each represented by a different DAG. We doubt that such mixtures generally exhibit the conditional independencies represented by a graphical chain model with undirected edges, but in any case the SDLC suggestion remains to be demonstrated. If feedback processes are represented by directed cyclic graphs, then it follows from LW that graphical chain models cannot represent them. Neither do graphical chain models represent the marginal conditional independence relations among observed variables that follow by the Markov condition from DAG models with latent variables (although other sorts of graphs that CW describe, but whose causal interpretation is not clear, can represent some marginal distributions of this kind). Graphical chain models could be used to represent a collection of alternative DAG models when one is unsure as to which structure is correct and the structures share certain conditional independence properties, but SDLC and CW and the papers they review do not unequivocally offer this interpretation.

The question of how the various "nondecomposable" forms of conditional independence relations described in CW could have been generated receives a straightforward answer using different generalizations of DAG models. Rather than starting with sets of conditional independence relations, finding a graphical formalism to represent them and then asking what causal process could have generated the constraints, we start with various sorts of causal processes represented by directed graphs and ask what sort of sets of conditional independence relations or marginal conditional independence relations they generate. It is important to be willing to abandon the idea, characteristic of graphical chain models, that the absence of an edge between two variables X and Y (which has a clear causal interpretation, namely that X does not directly cause Y) must always represent some conditional independence between X and Y; otherwise one excludes the natural representation of feedback processes. Two relevant generalizations of DAG models have been investigated.

1.2 Feedback and Reciprocal Causation

For many pairs of variables, A influences B and B influences A, whether directly or through some other set of variables considered in the system. Feedback processes can be represented by time series, but for linear systems they are often represented as well by finite directed cyclic graphs (DCGs). Methods for calculating correlations for cyclic systems flow from the

work of Haavelmo (1943) and Mason (1956). Despite this pedigree, even in the linear case very little is known about the connections between DCGs and conditional independence properties. The various conditions we have mentioned extend naturally to cyclic graphs, but the relationships among the conditions are different in cyclic and acyclic graphs. In the acyclic case, it is possible to define a graphical condition, *d*-separation, (Pearl, 1988) between three disjoint sets of variables **X**, **Y** and **Z** in a DAG G, such that **X** is *d*-separated from **Y** given **Z** if and only if the Markov condition applied to G entails that **X** is independent of **Y** given **Z**. In cyclic graphs the natural extension of the Markov condition does not capture all of the atomic independencies entailed by the natural extension of *d*-separation, and some formulations of the Markov condition are uninformative when extended to cyclic graphs (at least in the linear case).

In the case of linear normal models with unspecified values of some linear coefficients, there is a clear association of families of probability distributions with cyclic graphs, but we do not know in general how to characterize the conditional independence relations a linear normal cyclic system entails for all values of its free parameters. There is a purely graphical necessary and sufficient condition for a cyclic graph to require (for all linear models associated with it) that $\rho_{XY.U} = 0$, where U is a single variable (Glymour et al., 1987). The condition is in fact equivalent to a special case of *d*-separation for cyclic graphs. We have examined several four-variable cyclic graphs, and we find that the vanishing partial correlations of second order they require (again assuming linearity) also agree with the generalization of *d*-separation to cyclic graphs. There is no established convention for association of probability distributions with DCGs in the nonlinear case, but the linear case suggests that given the "right" association *d*-separation may correctly characterize the set of conditional independence relations common to all of the distributions associated with the graph. [Added in proof: The Markov condition in fact *fails* for some linear models (with correlated errors) for DCGs. For example, $x_3 = a\,x_1 + b\,x_4 + \varepsilon_3$ and $x_4 = c\,x_2 + d\,x_3 + \varepsilon_4$ does not entail that $P_{23.14} = 0$, as required by the Markov condition for the graph $x_1 \rightarrow x_3 \rightleftarrows x_4 \leftarrow x_2$. Spirtes has proven that *d*-separation does characterize the vanishing partial correlations implied by all linear models (with corrected errors) associated with any DCGs. See Directed Cyclic Graphs, Conditional Independence, Non-Recursive Linear Structural Equation Models, Carnegie Mellon Univ. Technical Report Phil-35, Dept. of Philosophy, 1993.]

1.3 Latent Variables

Consider a DAG G representing a causal process and any associated probability distribution P, where $\langle G,P\rangle$ satisfy Markov condition. Suppose that only a proper subset **O** of variables in the graph are measured or recorded. What conditional independence relation among variables in **O** is required by the Markov condition applied to G? What graphical object represents those marginal conditional independence relations and also represents information about G? A nice answer to both questions is given in Verma and Pearl (1990). They introduce the notion of the *inducing path graph* for G which contains only measured variables in G, encodes all of the marginal conditional independence relations G entails (by the Markov condition) and includes some of the causal information represented in G.

An undirected path U between X and Y is an *inducing path* over **O** in G if and only if (i) every member of **O** on U except for the endpoints occurs at the collision of two arrowheads on the path, and (ii) for every vertex V on U where two arrowheads collide, there is a directed path from V to X or from V to Y. There is an inducing path between X and Y in G over **O** if and only if X and Y are not independent conditional on any subset of $\mathbf{O}\backslash\{X,Y\}$. For variables X, Y in **O**, in the inducing path graph H for G over **O**, $X \leftrightarrow Y$ in H if and only if there is an inducing path between X and Y over **O** in G that is directed into X and also directed into Y; there is an edge $X \rightarrow Y$ in G if and only if there is no edge $X \leftrightarrow Y$ in H, and there is an inducing path between X and Y over **O** in G that is out of X and into Y. (It is easy to show that there are no inducing paths connecting X, Y in G over **O** that are not directed into X or into Y.) The two kinds of edges in an inducing path graph H have a straightforward causal interpretation: A directed edge $X \rightarrow Y$ occurs in H only if there is a directed path from X to Y in G, that is, X is a cause of Y; a double-headed edge $X \leftrightarrow Y$ occurs in H only if there is an unmeasured T and a directed path from T to X and a directed path from T to Y, the two paths intersecting only at T, that is, only if X and Y have an unmeasured common cause.

Unfortunately, observed conditional independence relations do not generally determine a unique inducing path graph, and so both for the purpose of studying causal inference and for characterizing indistinguishability of latent variable DAG models, another structure is required. A *partially oriented inducing path graph* (or POIPG for brevity) over a subset of variables **O**, represents a class of inducing path graphs over **O** that share the same adjacencies. A POIPG looks like an inducing path graph, but with the presence or absence of some arrowheads left unspecified. A directed edge in a POIPG indicates that all inducing path graphs in the class have that edge; a bidirected edge indicates that all inducing path graphs in the class have that bidirected edge. POIPGs can have edges ending in a mark, an "o," as in X o$\rightarrow$ Y, allowing some of the inducing path graphs represented to have $X \leftrightarrow$

Y and some to have $X \rightarrow Y$. Similarly, a POIPG may contain an edge X o-o Y. Two edges sharing a vertex, each with a mark at that vertex, can be underlined, as in -o $\underline{X}$ o-, indicating that the two "o" marks cannot simultaneously be arrowheads in any inducing path graph it represents. For some latent variable causal structures and sets of measured variables, the hypothesis that one measured variable does (or does not) cause another measured variable, or that two measured variables are affected by a latent common cause, can be read from the POIPG constructed from the conditional independence relations among the measured variables.

Spirtes (1992) describes a procedure for constructing a POIPG from conditional independence relations among observed variables and optional background knowledge, and Spirtes and Verma (1992) adapt this result to provide a polynomial time procedure to decide indistinguishability (by conditional independence) of any two DAGs with latent variables, assuming the Markov condition. Three examples of POIPGs are given in Figure 1 (ii), (iii) and (vii).

The DCG models and the POIPGs provide representations of most of the nondecomposable sets of independence hypotheses discussed by CW and explain how such independence properties could be generated. Of the five nondecomposable sets of independence hypotheses CW describe, four can be generated by a feedback process or a process with unmeasured common causes and represented by a DCG or POIPG. The fifth set of nondecomposable independencies can be generated by a cyclic graph but only with special parameter values (i.e., unfaithfully). Referring to CW's eight cases:

(i) $Y \perp\!\!\!\perp W$ (X,V) and $X \perp\!\!\!\perp V \mid (Y,W)$: DCG with $Y \rightarrow X \rightarrow W \rightarrow V \rightarrow Y$ or all arrows reversed [represented by the cyclic graph in Figure 1 (i)].

(ii) $Y \perp\!\!\!\perp W \mid V$ and $X \perp\!\!\!\perp V \mid W$ [represented by the POIPG in Figure 1 (ii)].

(iii) $Y \perp\!\!\!\perp W$ and $X \perp\!\!\!\perp V$: [represented by the POIPG in Figure 1 (iii)].

(iv) (v) and (vi) are represented by DAGs.

(vii) $Y \perp\!\!\!\perp W$ and $X \perp\!\!\!\perp V$ and $V \perp\!\!\!\perp W$ [represented by the POIPG in Figure 1 (vii)]. The POIPG in Figure 1 (vii) actually represents these independence relations only under the assumption of composition; that is, that for any four disjoint sets of random variables, **X, Y, Z, W**, the relations $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ and $\mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z}$ entail $\mathbf{X} \perp\!\!\!\perp (\mathbf{Y},\mathbf{W}) \mid \mathbf{Z}$. Composition holds for normal distributions.

(viii) $Y \perp\!\!\!\perp W \mid (X,V)$, $X \perp\!\!\!\perp V \mid (Y,W)$ and $V \perp\!\!\!\perp W$. This set of conditional and unconditional independence relations is not represented exactly by any DCG or POIPG unless the Faithfulness condition is violated.

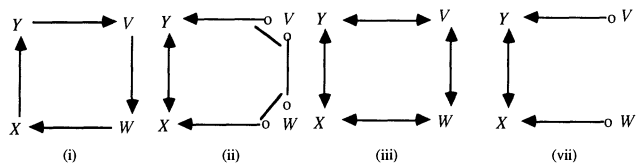

FIG. 1.

Most of the empirical examples CW give have small sample sizes, and the independence decisions are informal. In assessing the value of DCG and POIPG representations, it does not therefore seem important to consider whether feedback or latent variables are in these particular cases likely to be the correct substantive interpretations of the conditional independence relations.

2. CLASSIFICATION AND BAYESIAN NETWORKS

The construction of a Bayesian network expert system can be expensive and time consuming. Why bother? One use we can imagine is as a kind of personal calculator, a device an expert—or anyone who wishes to defer to and emulate that expert—can use to find out what her degrees of belief ought to be given various pieces of evidence. The expert, or expert emulator, can then use that information however she chooses in making decisions. In some contexts this seems to us a perfectly sensible purpose. Another conceivable purpose is to provide a system that combines prediction with explanations of how and why a prediction was obtained. Updating a Bayesian network resembles a course of reasoning, and perhaps some people may want such accounts of how predictions are obtained. But these are mostly advantages of computer-side manner. What advantages do Bayesian networks have as tools for furthering our knowledge and control of empirical domains?

Consider predictions (which we will refer to as classifictions) of a variable or variables **Y** using a set of variables **X** as predictors, for new individuals or samples drawn from a fixed distribution. There are a variety of automatic classification methods now available: neural networks, automatically constructed Bayesian networks, various forms of regression, automatically constructed decision trees and combinations of these (Shaffer, 1993). There are also a number of methods that rely on expert knowledge, such as hand-crafted decision trees and hand-crafted expert Bayesian networks. In such problems, there is a good deal of psychological evidence that computerized models of experts make better predictions in many domains than do the experts themselves, but so also do simple algorithmic prediction methods—for example, linear or logistic regression—when there is a relevant database. Do expert system Bayesian networks (or automatically constructed

Bayesian networks) have any advantages in reliability or computational ease over these other methods of classification, and if so, under what conditions?

Research has just begun on these questions, and the jury is still out on whether a Bayesian network constructed by consulting an expert makes superior classifications. SDLC note that all versions of the CHILD network with graphical structure extracted from an expert do less well at diagnostic prediction than does a "simple algorithmic" method (a hand-crafted decision tree). Moreover, SDLC compare the predictive accuracy of the network with fixed parameters estimated by the expert and with parameters changed by conditioning on data from new cases—unsurprisingly, the latter is superior—but they give no comparison with the predictive accuracy of the network when the parameters are estimated as much as possible entirely from the data. We wonder whether the model using parameter estimates based as much as possible on frequencies would (at least for some sample sizes) in this case do better than either of the methods of estimating parameters which they compare. The application of a Bayesian network constructed by consultation with an expert appears even more dubious in domains, such as psychology and sociology, in which rather less is known about causal mechanisms.

The graphical structure of Bayesian networks typically entails constraints on the joint distribution of measured variables. We expect a predictor that entails conditional independence constraints satisfied by the population distribution to have a smaller expected squared error than a predictor that does not, but the value of this advantage depends on our capacity to identify those constraints correctly: a predictor entailing a constraint false in the population will be biased. It seems to us a dicey question whether reductions in the variance of estimates are worth the risks of bias occasioned by assuming special conditional independence constraints on a distribution. Whatever the final result, it appears to us that while the method of constructing Bayesian networks with the aid of experts shows promise and is certainly worthy of further research, no decisive case has yet been made for the value of building Bayesian networks or causal models for the purpose of predicting within samples from a fixed distribution.

3. OTHER USES OF BAYESIAN NETWORKS AND CAUSAL MODELS

In the preceding section we used the qualifier "within a fixed distribution" because we believe the special value of DAG causal models is in predicting the results of interventions that change the distribution of variable values in a population. Predictions of this sort are not considered in the SDLC paper, but they are often the very point of causal models in studies that aim to influence policy. Such predictions can be made if one knows the causal structure of the systems in the population and understands the direct effects of the intervention. Unlike prediction within a fixed distribution, predictions of the outcomes of interventions absolutely require the use of the causal relations represented in the directed graph. Regression or other methods which take no account of causal structure will not suffice.

In a Bayesian network, given values for X on a new unit, we estimate the value of Y by computing the conditional probability of Y given X and doing whatever with the result. For a trivial example, suppose the network is Figure 2 (i) with binary variables, value 1 indicating the condition and 0 indicating its absence. The parameters of the network are $P(Smoking)$, $P(Yellow\ fingers \mid Smoking)$ and $P(Cancer \mid Smoking)$. If someone presents without yellow fingers we can compute $P(Cancer \mid Yellow\ fingers = 0)$; much of the SDLC review is devoted to how to perform such calculations in more complex cases. But what if, after constructing the network, we were to adopt a policy that prevents yellow fingers? Suppose we make everyone wash their hands twice a day and wear gloves in between, convenient gloves that do not make smoking more difficult and that are not carcinogenic. Assume our Bayesian network correctly describes the distribution of yellow fingers, smoking and cancer in the population before the new policy. Can the network be used to predict the probability of cancer in someone without yellow fingers after the policy is effected? Not by computing $P(Cancer \mid Yellow\ fingers = 0)$ as we did before. Instead we compute $P_{new}(Cancer \mid Yellow\ fingers = 0) = P_{new}(Cancer) = P(Cancer \mid Smoking)P(Smoking)$ (assuming after the policy is adopted no one has yellow fingers.) This is exactly the computation appropriate for the different network shown in Figure 2 (ii) with parameters $P_{new}(Yellow\ fingers)$, $P(Smoking)$, $P(Cancer \mid Smoking)$. The new network is obtained from the old by removing the directed edge from *Smoking* into *Yellow fingers*, giving *Yellow fingers* a new exogenous distribution and leaving the other parameters unchanged. The relation between the new network describing the distribution after the intervention and the original network describing the distribution before the intervention perfectly reflects the hypothetical facts: with the policy in place,

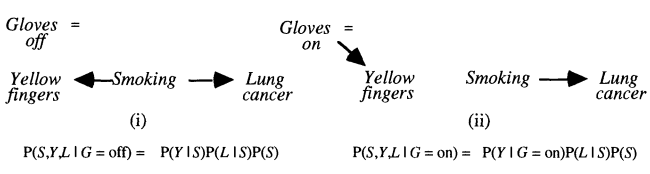

FIG. 2.

smoking no longer causes yellow fingers; the policy changes the probability of yellow fingers (to 0, or pretty close), but because yellow fingers do not cause either smoking or cancer, the new policy does not alter the joint distribution of these two variables.

Interventions to exogenously determine the distribution of values of a variable X, and that affect other variables only through X, break whatever edges into X originally obtained in the graph or graphs describing the causal structure(s) in the population and reparameterize the joint distribution accordingly. (Other kinds of interventions, which we do not consider here, may introduce as well as break edges.) We say X is "directly manipulated" by the intervention. The analysis is not ad hoc. When an intervention variable is introduced (*Gloves*, in the example) and the original distribution is understood to be conditional on a particular value of the intervention variables (e.g., *Gloves* = 0), the rule just illustrated follows from the Markov condition. A general proof is given in Spirtes, Glymour and Scheines, (1993).

This simple principle is at the center of experimental design. In graphical terms, Fisher wanted to randomize because he believed determining treatment by randomization guarantees that the structures describing the experiment will then contain no edges from causes of the outcome variable into the treatment variable. Rubin's proposals for causal inference in experimental designs (Rubin, 1974, 1977) and their extension by Pratt and Schlaifer (1988), are all consequences of the Markov condition for the special cases in which the intervention entirely determines the distribution of the variable or variables directly manipulated. The principle also explains features of distributions assumed in Bayesian discussions of experimental design (Kadane and Seidenfeld, 1992).

So it is easy in principle to determine the effects of a policy intervention provided one has a correct description of causal structure and a parameterization of the population distribution, and one knows the distribution of the directly manipulated variables that will result from the policy. Prediction of the outcomes of interventions is not so obvious if only a POIPG is available – and a POIPG is the best way we know of to characterize causal structure (without feedback) from observed conditional independence relations. There is, however, an algorithm that, given a POIPG and a set of measured variables to be directly manipulated, gives sufficient conditions and necessary conditions under which other variables can be predicted, and computes the new distribution (of a predictable variable) given the original joint distribution and the postpolicy distribution of the directly manipulated variables. (Spirtes and Glymour, 1993; Spirtes, Glymour and Scheines, 1993).

4. MODEL DISCOVERY

Extracting causal and probability information from experts can be time consuming and difficult even when the experts have real knowledge. Worse, in many problems the real knowledge of experts is quite limited, and according to a considerable psychological literature experts in many subjects know substantially less than they think they do. So we should be interested in fast, reliable procedures that can combine fragmentary prior knowledge with data to specify or partially specify causal or structural models. Few topics are more controversial in statistics, or, in our experience, more apt to draw scorn rather than research, although explicit arguments against the very idea (as opposed to arguments against particular procedures that have been proposed) tend to be feeble. For example, that "any data can be fit by several alternative models" (Rodgers and Maranto, 1989), or that there is no mechanical way to tell whether statistical dependencies are generated by an unknown causal process or by chance. Were the first objection sound a parallel would apply to all of statistical estimation. The second objection overlooks that humans can have some conviction that statistical dependencies are due to some causal process without knowing what that process is, and that even absent experimental manipulations, the very existence of a sensible model that explains puzzling features of a sample, may reasonably increase our conviction that the data are not a chance artifact.

Especially when it can be assumed that there are no latent factors at work, in our view directed graphical model specification is essentially a form of set valued estimation involving unfamiliar parameters, but subject to the same concerns for asymptotic reliability, error probabilities, variation of estimates and so on, as is ordinary parameter estimation. In the absence of strong prior information, model estimates should be set valued exactly because of indistinguishability classifications noted by SDLC. A classical version of the estimation theory should provide computable, consistent estimators; a Bayesian version should show how to compute at least the posterior mode and show that in the large sample limit the procedure yields the correct model – or class of models – almost surely.

A rudimentary theory of this kind already exists. SDLC note the results of Cooper and Herskovits (1992), which, given a linear ordering of discrete valued variables, for Dirichlet priors find the DAG compatible with the ordering and distribution that is the posterior mode on the sample evidence. Substituting a heuristic greedy search algorithm for the correct procedure, which is computationally intractable, their K2 algorithm is fast even for quite large numbers of variables and performs extremely well on simulated large sam-

ples. A non-Bayesian procedure, the PC algorithm, provably generates the set of all DAGs that represent (according to the Markov and Faithfulness conditions) a set of conditional independence facts in a distribution (assuming such a DAG exists). Prior ordering or partial ordering is optional, and the output may direct some or even all edges, depending on the structure of the correct DAG, even if no ordering information is input. The procedure minimizes the number of conditional independence tests required and the size of the set of variables conditioned on in each required test. PC has been implemented for multinormal and for multinomial variates in the TETRAD II program (Spirtes et al., forthcoming). The computational demands of the procedure depend on the sparseness of G. For fixed maximal degree, computation increases in the worst case as a polynomial function of the number of vertices. The procedure can be readily integrated with prior knowledge restricting G, and its error probabilities, as functions of sample size and average degree, have been investigated in extensive simulation studies with random graphs and randomly generated multinormal distributions. Wedelin (1993) has recently reported a procedure, so far implemented only for binary variables, that uses a parametrization related to the Fourier transform and an iterative algorithm for approximate maximum likelihood estimation of DAG models. The estimation is interleaved with an algorithm using Minimum Description Length criteria to construct a DAG, or an indistinguishability class of DAGs, from the data. The procedure is asymptotically correct for DAGs paired with faithful multinomial distributions. It does not require prior information about the ordering of the variables and has produced excellent results on simulated data with large numbers of variables.

SDLC briefly discuss the BIFROST program which generates chain graphs, described in more detail by Lauritzen, Thiesson and Spiegelhalter, (1992) (LTS) and illustrated again with data for the CHILD network. The program requires as input a partial ordering of the variables by blocks. It is not clear from this description whether the algorithm is practical for large numbers of variables, whether it is asymptotically correct, and to what extent the correct output depends on correctly specifying the block structure. We would like to know how the procedure performs on larger problems such as the ALARM network (Beinlich et al., 1989) for emergency medicine, which contains 37 variables, and has been used in tests of the reliability of the three procedures previously mentioned.

All of the algorithms so far described assume there are no latent common causes of measured variables. In real problems we often do not know at the outset whether statistical dependencies may be due to unmeasured factors affecting two or more measured variables. Absent some bound on the number of variables, there is an infinity of alternative DAGs that may accord with a set of observed conditional independence facts assuming the Markov and Faithfulness conditions, and there is no possibility of estimating a finite indistinguishability class of DAGs. What might be wanted instead are inference procedures that will describe features common to all DAGs admitting distributions yielding features of the observed marginal distribution, that is, POIPGs. It is often suggested that absent experimental interventions these kinds of inference cannot be correctly made even in principle, but with reasonable background assumptions that is not true. A correct algorithm for inferring POIPGs from conditional independence relations among observed variables is the FCI procedure given in Spirtes (1992), whose output is a POIPG. The procedure has been implemented for multinomial and multinormal distributions. The Spirtes and Verma algorithm, noted earlier, for deciding indistinguishability(by conditional independence) of DAGs with unobserved variables depends on the fact that POIPGs obtained by the FCI algorithm completely characterize the observed marginal conditional independence constraints entailed for the subset of observed variables by a DAG with latent variables. The procedure recovers each of the POIPGS (ii), (iii) and (vii) in Figure 1 from the corresponding conditional independence relations CW provide and also the undirected version of the cyclic graph in (i) (although we have no general proof that the algorithm correctly recovers cyclic graphs), as well as much more complicated structures in other cases.