A Review of Automated Procedures for the Discovery of Causal Relations and the Use of Causal Hypotheses in Prediction

Clark Glymour
Carnegie Mellon University

Abstract

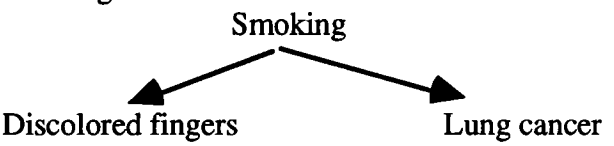
1. Motivation. The discovery tasks that have been best studied in statistical and computer science concern the extraction of statistical or rule based regularities from various sorts of data streams. In the natural and social sciences, however, practitioners want (and often claim) insight into the causal dependencies and mechanisms that produce observed patterns and rule instantiations. In policy contexts information is wanted that will enable correct prediction about relevant outcomes of interventions or actions in a system or population. Both epidemiology and economics are concerned with what will happen if alternative possible policies are broadly applied; the same is true of businesses, governments and many other organizations. Statistical or rule based patterns do not directly give any such information. For example, lung cancer is correlated with a history of slightly discolored fingers, but many interventions to prevent discoloration--requiring everyone to wear gloves, for instance--will have no effect on lung cancer or on lung cancer rates in the population so treated. In many contexts causal knowledge is limited, incomplete, or simply erroneous. The standard methods to extend causal knowledge involve experimental procedures in which various conditions are deliberately manipulated. But in many contexts--astronomy, geology, meteorology, epidemiology, human affairs, parts of engineering, parts of biology, and occasionally even in chemistry and physics--the requisite experiments cannot (or cannot all) be performed for practical or economic or ethical reasons. So what is wanted are reliable methods for combining limited prior knowledge with non-experimental (or partially non-experimental) data to extend causal knowledge or at least to guide us in deciding which experiments are worth performing; in other words, reliable methods are wanted for extracting causal explanations of observed statistical or rule based patterns. In addition, algorithms are wanted for predicting, whenever possible, the outcomes of interventions given partial and possibly quite incomplete causal knowledge and associate partial specification of probabilities.

In the face of widespread and dogmatic opinion in the statistical community that such procedures for discovery and prediction are impossible, the tasks of characterizing exactly the conditions for their possibility, and of finding and implementing appropriate algorithms, has fallen largely to the computer science community (if you are charitable enough to include in that group several of my colleagues in the Philosophy Department at Carnegie Mellon), with occasional help from exceptional statisticians, for example S. Lauritzen, N. Wermuth, and S. Fienberg. This talk will briefly survey some of the revolutionary progress in this subject in the last eight years, and will touch upon a number of open problems.

2. Sources. The notion that correlations, or more broadly, statistical dependencies found in nature or society are somehow to be accounted for by causal explanations is as old as modern science. In the 17th century, Francis Bacon described a semi-algorithmic method--the ancestor of modern concept learning systems--for extracting causal relations, and in the 19th century his procedure was repeated by Mill in his famous Methods for causal inference. In the 18th century causal inference was the focus of D. Hume's restatement of ancient skepticism. Bayes' famous paper was introduced and appended (by R. Price) as a refutation of Hume's skepticism about the possibility of reliable causal inference from statistical dependencies. In the 19th century G. Boole developed his logical theory as a preliminary to a larger project in which probability was to be used to infer causal relations, and late in that same century the notion of correlation was introduced as a measure of statistical dependency in normally distributed populations and as an indicator of underlying causal processes. In 1904, C. Spearman introduced into psychology theories or "models" in which measured variables are said to be linear functions of a common unmeasured variable and of uncorrelated, unmeasured noise variables specific to each measured variable; Spearman used his models to explain correlations found among psychometric variables. Subsequently, Yule's work made popular the idea of partial

correlation, which for normally distributed variables is a measure of statistical dependence conditional on one or more variables; Yule also pointed out especially problematic aspects to the causal interpretation of correlations in time series. In Spearman's models, and in those made popular subsequently by Thurstone's factor analysis, correlated measured variables have vanishing partial correlations when the common unmeasured variables are conditioned on. This and related work led to attempts at explicit analysis of the connections between probability and causality. H. Reichenbach proposed that observed statistical dependencies are to be explained by unmeasured common causes; later, H. Simon made a similar but more modest proposal, and explored the idea that in appropriately complete systems of variables causal relations can be determined by identifiability relations among variables. In 1970, P. Suppes proposed that "event A causes event B" just if A precedes B and there is no distinct third event C prior to B conditional on which A and B are independent. Although in the 1920s S. Wright introduced directed graphs to represent causal hypotheses in the context of statistical models, and they had been frequently used in that way thereafter, a general formulation of the intuitions connecting probability and causality using directed graphs to represent causal hypotheses only appeared in 1982 in the form of the Markov condition proposed by Kiiveri and Speed. Variants of their analysis were soon given by Lauritzen and Wermuth. In a series of papers culminating in an important book, J. Pearl explored the axiomatization of conditional independencies implied by the Markov condition, introduced additional important constraints connecting directed graphs and probability distributions, described algorithms for learning a simple class of graphs from probability distributions, and explored the use of Bayes networks in non-monotonic reasoning. More general search algorithms, some with proofs of their asymptotic correctness and completeness, have since been developed (individually or through collaborations) by G. Cooper, D. Geiger, myself, D. Heckerman, T. Richardson, P. Spirtes, R. Scheines, Pearl, T. Verma, D. Wedelin, and perhaps others I do not know of. An algorithmic analysis of prediction with a special form of partial causal knowledge represented as an event tree was given in 1986 by J. Robins, an epidemiologist, and independently discovered and formulated for directed graphs by Spirtes, Glymour, Scheines, Meek, Fienberg and Slate in 1991. Spirtes and Glymour presented an algorithm that characterizes predictions possible when causal information is extracted from a variable set (by the FCI algorithm--see below) that may exhibit statistical dependencies due to unmeasured and unknown common causes, and Pearl has investigated predictions when partial knowledge of unrecorded common causes is available. A general survey, up to date as of the end of 1992, is available in Spirtes, Glymour and Scheines, 1993. Much has happened since.

3. Conditioning and Intervention: Belief Networks and Causal Networks. The following network



can be viewed as simply a representation of constraints on the joint probability distribution of {Smoking, Discolored fingers, Lung Cancer}, where each feature is viewed as a binary variable. So understood, the graph asserts that Discolored fingers and Lung Cancer are not independent, but are independent conditional on Smoking; or, equivalently, that the joint distribution can be factored as $P(D,S,L) = P(D|S)P(L|S)P(S)$. Under a somewhat strengthened but still purely probabilistic interpretation, the graph also represents the claim that no other independencies or conditional independencies hold, implying for example that the probability of Lung cancer conditional on Discolored fingers is different from the unconditional probability of Lung cancer. Alternatively the graph can be viewed as encoding both the statistical constraints and also hypotheses about the statistical relations that will obtain if various interventions are carried out. In the causal interpretation, the graph represents the claims, for example, that a direct intervention to discolor fingers will not alter the probability of Smoking or the probability of Lung Cancer, but a direct intervention to alter smoking will give Lung cancer the corresponding conditional probability.

Belief networks and causal networks can share several assumptions, including the Markov condition introduced by Kiiveri and Speed and the Minimality and Faithfulness conditions introduced (the last in other terms) by Pearl. It

must be emphasized that these are restrictions on *pairings* of directed graphs and probability distributions:

Markov: If G is a directed graph with vertex set V then for all distinct X, Y in V, if X is not a descendant of Y, then X is independent of Y conditional on the parents of Y.

Minimality: If G is a directed graph, a probability distribution satisfying the Markov condition for G does not satisfy the Markov condition for any proper subgraph of G.

Faithfulness: If G is a directed graph, and X, Y are distinct members of V and Z is a subset of V not containing X or Y, then X Y are independent conditional on Z only if the Markov condition applied to G entails that independence.

The causal interpretation differs in applying the Markov condition only to sets of variables V such that every variable that directly influences two or more members of V is a member of V. Such sets will hereafter be called **causally sufficient**. In the example previously given, the Markov condition says that Discolored fingers and Lung cancer are independent conditional on Smoking. The Minimality condition says, for example, that Smoking and Discolored fingers are not independent--for if they were the distribution would satisfy the Markov condition for the subgraph in which the Smoking -> Discolored fingers edge is removed. The Faithfulness conditions says, for example, that Smoking and Discolored fingers are not independent and are not independent conditional on Lung cancer. Markov + Faithfulness implies Minimality, but Markov + Minimality does not imply Faithfulness. Glymour and Spirtes proved that for linear systems, using a natural measure on the parameters, the set of distributions violating Faithfulness for any given graph has measure zero; a similar argument has been given by Meek for discretely valued variables. Small sample distributions, however, often appear to violate Faithfulness, as can systems in which some variables are deterministic functions of other variables.

In using the Markov condition, the essential fact is that it implies that a probability distribution associated with a directed graph can be factored into a product of conditional and unconditional probabilities. Start with the terminal vertices and write their respective probabilities conditional on their parents; do the same with each of their parents, etc., and write the probabilities of zero indegree nodes (i.e., those without parents); the product of these probabilities must equal the joint distribution if the Markov condition is satisfied.

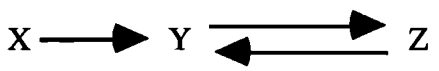
Pearl, Geiger and Verma gave a polynomial time algorithm--the d-separation procedure-- for deciding whether the Markov condition implies, for any graph, any specified conditional independence. Lauritzen provided an alternative, equivalent algorithm which determines whether the marginal density over X, Y and Z and their ancestors in the graph can be factored into two parts sharing only Z. Geiger gave a complete characterization of the conditional independence relations implied when child variables are deterministic functions of their parents, and Spirtes gave an entirely general characterization of independence with any deterministic relations. A completeness proof for Spirtes' characterization has not yet been provided.

4. Statistical Indistinguishability. Graphs that can represent exactly the same set of probability distributions do not, under a causal interpretation, say the same thing, but they cannot be distinguished by any inferences based on sample frequencies; they must be distinguished, if at all, either by prior knowledge or experimental intervention or by imbedding them in a larger set of causal relations. Thus under any of the axioms of the previous section, Smoking -> Lung cancer and Smoking <- Lung cancer and Smoking <- Genotype -> Lung cancer represent the same families of probability distributions over {Smoking, Lung cancer}, and could not be distinguished by sample frequencies alone, although what the three hypotheses assert is importantly different. Indistinguishability characterizations provide the criteria of adequacy for causal discovery procedures: in the large sample limit, the procedures should, in the absence of prior knowledge, give the appropriate indistingusihability class. Further, indistinguishability algorithms, when combined with statistical tests, often immediately lead to discovery procedures that are correct in the large sample limit, but which may not be computationally efficient. The first theoretical problem in considering automated discovery of causal structure is therefore to characterize the indistinguishability classes of graphs under various assumptions about the relations between graphs and probabilities. Frydenberg, and independently Verma and Pearl, showed that for

directed acyclic graphs the set of distributions satisfying the Markov condition for two graphs are the same if and only if (i) the graphs have the same vertex set; (ii) the same adjacencies; (iii) for all vertices, X,Y,Z, with X, Z not adjacent, X ->Y <- Z in one graph if and only if in the other. The same condition characterizes indistinguishability assuming both Markov and Faithfulness. Spirtes showed that indistinguishability with the Markov and Minimality conditions holds if and only if (i) and (ii) above hold and for all vertices X, Y, Z, X -> Y <- Z in one graph if and only if in the other. Under any of these sets of assumptions, any two graphs on the same vertex set can be extended to distinguishable graphs by adding the same additional variables and the same edges from new variables to old.

The most common objection to causal inference from uncontrolled data is that observed statistical dependencies may be due to unrecorded common causes, and any principled study of inference therefore requires consideration of that possibility. Users of linear models often specify correlations among exogenous variables, in apparent violation of the Markov condition; such systems can, however, always be made consistent with the Markov principle for correlations by replacing each exogenous correlation with an unmeasured common cause of the correlated variables. The question of indistinguishability for structures with unmeasured or unrecorded variables is this: When do two directed acyclic graphs admit the same class of marginal probability distributions on a proper, common subset of their vertices, where the marginal is taken in each case from a joint distribution satisfying the Markov condition for the full graph? T. Verma developed a graphical characterization of the independence and conditional independence relations a given graph entails for any subset O of its vertex set. Verma's *inducing path graphs* contain only the observed vertices, but unlike directed graphs may contain double headed arrows. A directed edge X -> Y in an inducing path graph indicate the presence in the full graph, with unmeasured variables, of a structure (an "inducing path") that introduces a dependency between X and Y that cannot be eliminated by conditioning on any subset of O. An infinity of distinct directed graphs with extra vertices can nonetheless have the same inducing path graph. Spirtes introduced a the idea of a *partially ordered inducing path graph*, in which the direction of some edges are left unspecified, and devised a procedure, the FCI algorithm, to determine a partially oriented inducing path graph from observed independencies and conditional independence's. Spirtes and Verma proved that two directed acyclic graphs are indistinguishable from conditional independencies on O if and only if the FCI algorithm produces the same partially oriented inducing path graph for the two structures. Indistinguishability can be solved in time polynomial in the number of vertices.

In all of the preceding it has been assumed that graphs considered are acyclic. In engineering and economic applications, however, feedback systems are often postulated that can be described by infinite time series. In both subjects the limiting ("equilibrium") distributions in time series are often described by systems of equations that determine directed *cyclic* graphs. For linear systems, the graphs correspond to sets of simultaneous equations in which each variable is written as a linear function of its parents plus a specific noise, and the noise or "error" variables are independently distributed. For linear systems corresponding to directed cyclic graphs, the Markov condition does not correctly capture conditional independence, as the following example illustrates:



X is not a descendant of Z and the only parent of Z is Y, but X and Z are not independent of Y conditional on Z, as may be verified by computing the partial correlation in the corresponding system of linear equations. Spirtes showed that for linear systems d-separation (or Lauritzen's algorithm) characterizes the conditional independence relations implied by a cyclic graph. Even d-separation fails for non-linear directed cyclic graphs however. In some of the most intricate mathematical work yet to have appeared in the subject, T. Richardson has given necessary and sufficient conditions for the indistinguishability within linear systems of any two directed cyclic graphs and has provided and implemented a polynomial time algorithm for deciding equivalence. The work shows that while many cyclic graphs are equivalent to acyclic structures, there are cyclic graphs that are equivalent to no acyclic graph, not even with unmeasured variables.

5. Search Procedures, Reliability and Complexity. Four approaches to automated construction of causal networks have been investigated: (i) procedures that call for statistical tests of independence or conditional independence hypotheses; (ii) procedures that put a prior distribution over graphical features, posit a likelihood function for graphs and data, and update by Bayes rule or some heuristic approximation; (iii) procedures that use minimum description length methods; and (iv), combined procedures. The first approach is thus far the most developed, and algorithms using this approach are available commercially in the TETRAD II program. (i) Assuming that all common causes of measured variables are measured (causal sufficiency), the PC algorithm (Spirtes, et al. 1993) provides a fast feasible procedure for any application where statistical decisions can be made about conditional independence of variables. The algorithm, which is worst case exponential in the maximal degree of the true graph but polynomial (for fixed maximal degree) in the number of vertices, has been implemented in the TETRAD II program for normal and multinomial distributions, and has been extensively tested on simulated data. Dropping the assumption of causal sufficiency, the FCI algorithm (Spirtes et al, 1993) provides correct information about the DAGs even with unmeasured variables, that can generate conditional independence relations in the marginal distribution over measured variables. The algorithm is worst case exponential in the number of vertices even with fixed maximal degree. Neither the PC algorithm nor the FCI algorithm give the maximal possible information about orientation of directed edges; Meek (1995) and independently Madigan (1995) have proved complete a set of orientation rules (previously suggested by Verma) for the causally sufficient case. For normal variates and prior information clustering variables into sets whose respective members share *at least* one common unmeasured cause, procedures have been found to form from each cluster a subset of variables sharing only a single unmeasured common cause and no other causal relations, if such a subset exists. (Spirtes, et al. 1993). From such a purified collection of clusters of measured variables, a polynomial time (in the number of latent variables) has been found to determine the graphical structure of causal relations among the unobserved variables. These various procedures have been used in a variety of applications, including plant genetics, development of scales for measuring pain and for psychiatric variables, analysis of physical spectra, personnel studies, etc.(ii) The K2 algorithm of Cooper and Herskovits requires a prior linear ordering of discrete variables and a special prior distribution and uses a greedy heuristic search to output an estimate of the graph that is the posterior mode. The procedure is comparable in complexity to the PC algorithm, but seems to work better when a linear ordering of variables is available. Heckerman, et al., (1994) have developed Bayes scores (for both normal and multinomial distributions) that give the same score to statistically indistinguishable graphs, and have used the scores in search procedures that do not require a prior ordering but does require an initial graph. The procedures appear not to work very well unless the initial graph is very close to the true graph. No general Bayes procedures have been developed for non-causally sufficient cases. (iii) Wedelin (1993) has developed a minimum description length algorithm for learning DAGs in the causally sufficient case with binary variables. (iv) Recently Spirtes (1995) has considerably improved the performance of the PC algorithm by using the Heckerman, et al. Bayes scores in a post processor, checking each possible addition, deletion or reorientation of edges in the PC output. For the ALARM network (Beinlich, et al, 1987), for which the K2 algorithm requires a complete ordering of variables as prior information, and the Heckerman, et al procedures make a considerable number of errors even when given a nearly correct and complete graph as prior information, and the PC makes several errors with no prior information, Spirtes procedures reproduces the true graph with a single error, using no prior information. The output has a higher posterior probability than does the true model.

6. New Issues. There are a wide range of open, fundamental questions in the area. Purely statistical questions are open about simultaneous decisions about conditional independence in many distributions. Nothing is known about how much of the analysis of latent structure for normal variates can be extended to other distributions. Verma has shown that there are non-independence constraints among measured variables that result from some latent structures, but they have not been generally characterized or

put to use in discovery. No method is know of combining the FCI algorithm with Bayes procedures for the latent case, because the latter do not yet exist. Herron (1995) has given a counting principle for the number of DAGs in an equivalence class assuming Minimality, but none is available for the Markov condition alone. Wermuth has emphasized, and Cooper has vividly illustrated, that the Markov condition can fail if the sample is inappropriately selected, and he and the Carnegie Mellon group are at work characterizing the limits of reliable inference when there is sample selection bias. Pearl has investigated a range of questions about prediction with partial knowledge of causal structure, and many related questions remain open. One of the most interesting issues concerns feedback systems represented by directed cyclic graphs, for which the Markov condition fails but for which (in the linear case) Spirtes (1994) has shown that d-separation still characterizes conditional independence. Richardson (1995) is investigating search procedures for cyclic graphs and investigating the very interesting question of the relation between such graphs and linear time series. Psychological aspects of causal inference are virtually unexplored in all but trivial cases, and experiments to inquire into human ability to make reliable causal inference from statistics in multivariate cases are being conducted by Glymour and M. Pimm-Smith.

References: See P. Spirtes, C. Glymour and R. Scheines, Causation, Prediction and Search, Springer, 1993 for references until 1993; for other references see the bi-annual proceedings on AI and Statistics, and technical reports in Logic and Computation form the Department of Philosophy Carnegie Mellon University, from the Center for Cognitive Science, Department of Computer Science, University of California at Los Angeles, and from the Microsoft Decision Analysis group.