Data Mining and Knowledge Discovery 1, 203–224 (1997)

A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships

GREGORY F. COOPER
Center for Biomedical Informatics, Suite 8084, Forbes Tower, University of Pittsburgh, Pittsburgh, PA 15213

Editor:

Received September 17, 1996; Revised March 17, 1997; Accepted March 25, 1997

Abstract. This paper presents a simple, efficient computer-based method for discovering causal relationships from databases that contain observational data. Observational data is passively observed, as contrasted with experimental data. Most of the databases available for data mining are observational. There is great potential for mining such databases to discover causal relationships. We illustrate how observational data can constrain the causal relationships among measured variables, sometimes to the point that we can conclude that one variable is causing another variable. The presentation here is based on a constraint-based approach to causal discovery. A primary purpose of this paper is to present the constraint-based causal discovery method in the simplest possible fashion in order to (1) readily convey the basic ideas that underlie more complex constraint-based causal discovery techniques, and (2) permit interested readers to rapidly program and apply the method to their own databases, as a start toward using more elaborate causal discovery algorithms.

Keywords: causal discovery, data mining, observational data

1. Introduction

This paper presents a simple, efficient computer-based method for discovering (under assumptions) causal relationships from observational databases. A primary purpose for applying the method is to gain insight into the causal relationships among a set of database variables. Such knowledge permits one to know how varying a causal variable is likely to induce a change in an effect variable. For example, suppose we have a large database of information about thousands of patients seen in a hospital for the past five years. What causal relationships are suggested as highly likely by this observational data? If some of these causal relationships were previously unknown and are likely to have significant clinical benefit, then we may wish to follow up with additional studies and analyses.

Observational data is passively observed, as contrasted with experimental data in which one or more variables is manipulated (often randomly) and the effects on other variables are measured. Observational data is more readily available than experimental data, and indeed, most databases that are used for data mining are observational databases. As observational databases become increasingly available, the opportunities for causal discovery increase. Techniques for causal discovery could be applied in performing exploratory data analyses

of large databases of many different kinds. Such analyses might uncover, for example, how a variant in operating procedures is likely to influence productivity, or how using a particular marketing strategy is likely to change product sales. The potential application areas are wide ranging.

Traditional statistical thinking says that "correlation does not imply causation". Observational data can, however, constrain the causal relationships among variables. Perhaps the simplest example of such a constraint is the inductive principle that if two variables X and Y are not correlated (or, more generally, are not statistically dependent according to some measure), then X does not cause Y, and Y does not cause X. While this principle can fail, it also can serve as a powerful guide in the search for causal relationships. The story, however, as we relate it in part here, is much richer and more interesting than that simple principle. In particular, the most important general idea in this paper is that *information about the statistical independence and dependence relationships among a set of variables can be used to constrain* (*sometimes significantly*) *the possible causal relationships among a* subset *of those variables*. For example, suppose that in fact X causes Y. By measuring just X and Y, we indeed cannot determine whether X causes Y. So, in that limited sense, correlation does not imply causation. If, however, there is a variable W that is known not to be caused by X or Y, then by examining the statistical independence and dependence relationships among W, X, and Y, it sometimes is possible to infer that X causes Y. Section 3 shows how. In some instances, even though we may not be able to induce that X causes Y, we may be able to determine, for example, that Y does not cause X, and thereby constrain the possible causal relationships between X and Y.

This paper focuses on predictive causal relationships that express how changing X is likely to change Y. The paper does not address the meaning of causality from a philosophical perspective. Operationally, however, when we say that X *causes* Y we mean that a hypothetical, ideal randomized controlled experiment would conclude that there is some manipulation of X that leads to a change in the probability distribution of values that Y will take on.

We outline here a prototypical randomized controlled experiment (RCE); although variations certainly exist, they are not discussed. An RCE is performed with an explicitly defined population of units (e.g., patients with *chest pain*) in some explicitly defined context or set of contexts (e.g., currently receiving no chest-pain medication and residing in a given geographical area). Thus, causal relationships that are discovered are relative to a population and a context. In an RCE, for a given experimental unit, the value to set the cause in question, which we denote as X, is randomly selected using a uniform distribution over the other possible values of X. The state of X is then manipulated to have the selected value. The RCE defines explicitly the details of how these value manipulations are made (e.g., the quantity of chest-pain medication to take and the time course of taking the medication). For each unit, after the new value of X is set (e.g., either *receive chest-pain medication* or *receive no chest-pain medication*), the value of Y is measured (e.g., *either has chest pains* or *does not have chest pains*). The greater the experimental data support a statistical dependency between X and Y, the more the data support that X causally influences Y.

In practice, of course, even a limited randomized controlled experiment might not be safe, ethical, logistically feasible, financially worthwhile, or even theoretically possible, all of which are reasons for using observational data to attempt to infer causal relationships.

A primary purpose of this paper is to present a constraint-based causal discovery method in the simplest possible fashion; therefore, the coverage here is relatively informal and non-technical. By constraint-based, we mean a two-step procedure in which (1) statistical tests are used to establish conditional dependence and independence relationships among the variables in a model, and (2) those relationships are used to constrain the types of causal relationships that exist among the model variables. This presentation leads to an algorithm that is more efficient (in the worst case), but less complete, than previously published constraint-based causal discovery algorithms. Thus, the algorithm introduced here is relatively fast, but it often will not output all the causal relationships that might be discovered by more computationally complex algorithms. After reading this paper, readers familiar with computer programming should be able to implement in a matter of a few hours the discovery algorithm that is described. While the algorithm may not perform as well as more complex algorithms that are referenced in the paper, the algorithm here should provide a good starting point for using or even implementing those more sophisticated and complex algorithms. Additionally, the basic ideas presented here should make it easier for readers to understand the general theory of constraint-based causal discovery (Pearl and Verma, 1991; Spirtes et al., 1993). In Section 5, we discuss the relationship between constraint-based and Bayesian methods for causal discovery.

2. Assumptions for causal discovery

In this section, we describe six assumptions that are used to support the primary causal discovery algorithm presented in this paper.

Assumption 1 (Database completeness). Let D be a database of cases (e.g., a flat file of records), such that each case contains a value for each variable in set $\boldsymbol{V}$.

Assumption 2 (Discrete variables). Each variable in $\boldsymbol{V}$ has a finite, discrete number of possible values.

Assumption 2 is not required, but rather, it is made for convenience.

Assumption 3 (Bayesian network causal model). The underlying causal processes that exist among the variables in $\boldsymbol{V}$ can be modeled using some Bayesian network G, which might contain hidden variables not in $\boldsymbol{V}$.

Assumption 3 means in effect that we assume that the data we have about the variables in $\boldsymbol{V}$ were generated by some Bayesian network[1].

A Bayesian network consists of a structural model and a set of probabilities. The structural model is a directed acyclic graph in which nodes represent variables and arcs represent probabilistic dependence. For each node there is a probability distribution on that node given the state of its parents. A Bayesian network specifies graphically how the node probabilities factor to specify a joint probability distribution over all the nodes (variables). A causal Bayesian network is a Bayesian network in which the parents of a node are interpreted as directly causing that node, relative to the other nodes in the model. For a more detailed

discussion of Bayesian networks, see (Castillo et al., 1997; Jensen, 1996; Pearl, 1988; Spirtes et al., 1993).

Let S be the graphical structure of G and let P be the joint probability distribution represented by G. By definition, S is a directed, acyclic graph. A node in S denotes a variable that models a feature of a process, event, state, object, agent, etc., which we will denote generically as an entity. For example, age is a feature of a car, which is an object/entity. We will use the terms *variable* and *node* interchangeably. Also, as shorthand, we sometimes will say that one variable causes another variable, rather than say one variable represents a feature of an entity that causes a feature of another entity that is represented by another variable. Note that while S surely will contain all the variables in $\boldsymbol{V}$, it also may contain variables that are not in $\boldsymbol{V}$. In particular, we later will see that hidden variables may appear in S to explain the statistical dependence among the measured variables in $\boldsymbol{V}$.

An arc $X \rightarrow Y$ in S denotes direct causation of Y by X, relative to the other variables in S. Suppose, however, that there is some variable U, such that X only influences Y through U. We express this relationship in Bayesian network notation as $X \rightarrow U \rightarrow Y$. Here X is no longer a direct cause of Y (in $\boldsymbol{V}$), but rather, is an indirect cause.

Since we are using a Bayesian network model, the Markov condition must hold by definition. The *Markov condition* is as follows: Any node is conditionally independent of its nondescendants, given its parents. A nondescendant of a node X is a node Y that cannot be reached by a directed path from X to Y. The intuition underlying the causal version of the Markov condition is as follows. Assume that the structure S of a Bayesian network G is causally valid. A descendant Y of X in S is on a causal path from X. Thus, we would expect there to be the possibility of a probabilistic dependency between X and Y. Now, consider the nondescendants of X; that is, consider all entities represented by the variables in G that are not directly or indirectly caused by X. Since the parents of X represent all of its direct causes, if we fix the values of these parents, we expect that the nondescendants of X will be probabilistically independent of X, unless a nondescendant happens also to be on an effect of X; thus, they will give us no information about the distribution of X.

A criterion called *d-separation* captures exactly the conditional independence relationships that are implied by the Markov condition (Geiger et al., 1990; Meek, 1995; Pearl, 1988)[2]. The following is a definition of d-separation (Pearl, 1994): Let A, B, and C be disjoint subsets of the nodes in S. Let p be any acyclic path between a node in A and a node in B, where an acyclic path is any succession of arcs, regardless of their directions, such that no node along those arcs appears more than once. We say a node w has converging arrows along a path if two arcs on the path point to w. Subset C is said to block p if there is a node w on p satisfying one of the following two conditions: (1) w has converging arrows (along p) and neither w nor any of its descendants are in C, or (2) w does not have converging arrows (along p) and w is in C. Subset C is said to d-separate A from B in S if and only if C blocks every path from a node in A to a node in B.

We will use the d-separation condition to distinguish one causal model from another. In essence, it provides a link between the causal process that we do not perceive directly and the measurements that we do perceive. In order for this link to be tight, however, we need several more assumptions.

Assumption 4 (Causal faithfulness condition). For all disjoint sets A, B, and C in $\boldsymbol{V}$, if in S we have that A is not d-separated from B by C, then in P we have that A and B are conditionally dependent given C.

Assumption 4 says that the only way variables will be probabilistically independent is if their independence is due to the Markov condition, or equivalently, to the d-separation condition.

The following result regarding the faithfulness condition has been proved for discrete (Meek, 1995) and for multivariate Gaussian (Spirtes et al., 1993) Bayesian networks. Consider any *smooth* distribution Q over the possible parameters in a Bayesian network. The parameters are just the probabilities represented in the network, which we denoted above as P. Now consider drawing a particular set of parameters from distribution Q. The results in (Meek, 1995; Spirtes et al., 1993) show that the probability of drawing a distribution that is not faithful is measure zero. These results do not mean that drawing such a distribution is impossible, but rather, under the assumption of a smooth distribution, such an outcome is exceedingly unlikely.

Assumption 5 (No selection bias). If $\boldsymbol{V}'$ denotes an arbitrary instantiation of all the variables in $\boldsymbol{V}$, then $\boldsymbol{V}'$ is sampled for inclusion in D with probability $\Pr(\boldsymbol{V}' \mid G)$, where G is the causal Bayesian network to be discovered.

If selection bias exists, then it could be that $\boldsymbol{V}'$ is sampled with a probability other than $\Pr(\boldsymbol{V}' \mid G)$. In Section 4 we return briefly to the issue of selection bias and show how the violation of Assumption 5 can be detected in principle.

Let T be a test used to determine conditional independence among sets of variables in $\boldsymbol{V}$, as for example, the chi-squared test.

Assumption 6 (Valid statistical testing). Consider the sets of variables A, B, and C in $\boldsymbol{V}$. If in P we have that A and B are conditionally dependent given C, then A and B are conditionally dependent given C according to test T applied to the data in D. Similarly, if in P we have that A and B are conditionally independent given C, then A and B are conditionally independent given C according to test T applied to the data in D.

Assumption 6 states that we can use test T to uncover the probabilistic dependence and independence relationships among the measured variables, as given by P. Note that T implicitly includes the value of any statistical significance threshold (e.g., an alpha level) that is required in applying the test.

3. A causal discovery algorithm

In this section, we present a simple causal discovery algorithm, which we call LCD (Local Causal Discovery). As stated in the introduction, the purpose of this presentation is to illustrate the basic ideas that underlie constraint-based causal discovery from observational data, as well as present an efficient algorithm that may be useful in practice for data mining. We first present the conceptual basis of the algorithm. Next, we present the algorithm itself as pseudocode, and we characterize its computational complexity.

3.1. The basis of LCD

Suppose we know (or, more likely, we are willing to assume) that the feature (of some entity) represented by variable W in $\boldsymbol{V}$ is not caused by any of the features (of entities) represented

by the other variables in $\boldsymbol{V}$. We might, for instance, assume this causal constraint based on scientific or commonsense principles that we are willing to believe, direct observation, or the results of a randomized controlled experiment. For example, in a clinical database, W might represent patient gender. As another example, the occurrence of W might temporally precede all the other measured variables in $\boldsymbol{V}$.

Assumption 7 (A known uncaused entity). There is a designated variable W in $\boldsymbol{V}$ that is not caused by any other variable in $\boldsymbol{V}$.

Consider three measured variables W, X, and Y in $\boldsymbol{V}$. Table 1 shows the 4 ways that we will model how W and X can be causally related. The variable H_{WX} is a hidden variable that is causally linking just W and X. A hidden (i.e., latent) variable is a variable about which we have no measurements. Since a set of hidden variables can always be modeled using one composite hidden variable, there is no loss of generality in considering only one hidden variable here.

In a fashion parallel to Table 1, Table 2 lists the 4 ways that we will model how W and Y can be related.

Finally, Table 3 lists the 6 ways we will model how X and Y can be related; there are 6 ways, rather than 4, because we include the possibility that Y can cause X.

Table 1. The 4 modeled causal relationships between W and X.

Label	Causal relationship
$WX1$	$W \quad X$
$WX2$	$W \rightarrow X$
$WX3$	H_{WX} $\swarrow \quad \searrow$ $W \quad X$
$WX4$	H_{WX} $\swarrow \quad \searrow$ $W \rightarrow X$

Table 2. The 4 modeled causal relationships between W and Y.

Label	Causal relationship
$WY1$	$W \quad Y$
$WY2$	$W \rightarrow Y$
$WY3$	H_{WY} $\swarrow \quad \searrow$ $W \quad Y$
$WY4$	H_{WY} $\swarrow \quad \searrow$ $W \rightarrow Y$

Table 3. The 6 modeled causal relationships between X and Y.

Label	Causal relationship
$XY1$	$X \quad Y$
$XY2$	$X \rightarrow Y$
$XY3$	$X \leftarrow Y$
$XY4$	H_{XY} $\swarrow \quad \searrow$ $X \quad Y$
$XY5$	H_{XY} $\swarrow \quad \searrow$ $X \rightarrow Y$
$XY6$	H_{XY} $\swarrow \quad \searrow$ $X \leftarrow Y$

The relationships in Tables 1 through 3 do not model all the possible ways that we may have arcs between measured variables and hidden variables, the ways we may have arcs among the hidden variables, or the ways in which two or three variables can share a common hidden variable. At the expense of completeness, we exclude consideration of some possibilities, in order to clearly and succinctly convey fundamental concepts. The general theory of constraint-based causal discovery (Spirtes et al., 1993) shows that considering these additional possibilities does not interfere with the causal distinctions we make in this paper.

The arcs among measured variables in Tables 1 through 3 are direct causal relationships relative only to the set of variables $\{W, X, Y\}$, rather than relative to the set $\boldsymbol{V}$ of all variables. Consider, for example, the relationship $W \rightarrow X$ given as $WX2$ in Table 1. It could be that there is some variable U in $\boldsymbol{V}$ such that U represents an intermediate causal step in the causal chain from W to X, namely $W \rightarrow U \rightarrow X$. Thus, relative to consideration of just the variables in $\{W, X, Y\}$, U is in a sense hidden.

Table 4 uses the labels from Tables 1 through 3 to list in column one all 96 possible causal models that follow from the 4 modeled relationships between W and X, the 4 modeled relationships between W and Y, and the 6 modeled relationships between X and Y. Column two contains three d-separation conditions for each causal graph. The notation $D(W, X)$ is used to represent that W and X are not d-separated (i.e., they are dependent) in G. Similarly, $D(Y, Z)$ means that Y and Z are dependent. The notation $I(W, Y \mid X)$ represents that W and Y are d-separated (i.e., they are independent) given X. A "+" in Table 4 indicates the presence of the designated relationship; a blank indicates its absence. If test T is a reliable indicator of independence, as we assume, then these three conditions can be determined from the data in D.

The patterns of dependence and independence displayed in Table 4 form equivalence classes among the 96 causal graphs that are induced by the three d-separation conditions. The members of a particularly important equivalence class are networks 18, 19, and 20 in Table 4, which are highlighted by underlining them. These causal graphs, which all have the pattern $+\,+\,+$, are shown in Table 5. Call this pattern the *positive pattern*. Note that in

Table 4. A listing of three d-separation conditions for 96 causal graphs (see text).

	Causal graph	$D(W, X)$	$D(X, Y)$	$I(W, Y \mid X)$
1.	$WX1\ WY1\ XY1$			+
2.	$WX2\ WY1\ XY1$	+		+
3.	$WX3\ WY1\ XY1$	+		+
4.	$WX4\ WY1\ XY1$	+		+
5.	$WX1\ WY2\ XY1$			
6.	$WX2\ WY2\ XY1$	+	+	
7.	$WX3\ WY2\ XY1$	+	+	
8.	$WX4\ WY2\ XY1$	+	+	
9.	$WX1\ WY3\ XY1$			
10.	$WX2\ WY3\ XY1$	+	+	
11.	$WX3\ WY3\ XY1$	+		
12.	$WX4\ WY3\ XY1$	+	+	
13.	$WX1\ WY4\ XY1$			
14.	$WX2\ WY4\ XY1$	+	+	
15.	$WX3\ WY4\ XY1$	+	+	
16.	$WX4\ WY4\ XY1$	+	+	
17.	$WX1\ WY1\ XY2$		+	+
18.	$\underline{WX2\ WY1\ XY2}$	+	+	+
19.	$\underline{WX3\ WY1\ XY2}$	+	+	+
20.	$\underline{WX4\ WY1\ XY2}$	+	+	+
21.	$WX1\ WY2\ XY2$		+	
22.	$WX2\ WY2\ XY2$	+	+	
23.	$WX3\ WY2\ XY2$	+	+	
24.	$WX4\ WY2\ XY2$	+	+	
25.	$WX1\ WY3\ XY2$		+	
26.	$WX2\ WY3\ XY2$	+	+	
27.	$WX3\ WY3\ XY2$	+	+	
28.	$WX4\ WY3\ XY2$	+	+	
29.	$WX1\ WY4\ XY2$		+	
30.	$WX2\ WY4\ XY2$	+	+	
31.	$WX3\ WY4\ XY2$	+	+	
32.	$WX4\ WY4\ XY2$	+	+	
33.	$WX1\ WY1\ XY3$		+	+
34.	$WX2\ WY1\ XY3$	+	+	
35.	$WX3\ WY1\ XY3$	+	+	
36.	$WX4\ WY1\ XY3$	+	+	
37.	$WX1\ WY2\ XY3$	+	+	
38.	$WX2\ WY2\ XY3$	+	+	

(*Continued on next page.*)

Table 4. (*Continued.*)

	Causal graph	$D(W, X)$	$D(X, Y)$	$I(W, Y \mid X)$
39.	$WX3\ WY2\ XY3$	+	+	
40.	$WX4\ WY2\ XY3$	+	+	
41.	$WX1\ WY3\ XY3$	+	+	
42.	$WX2\ WY3\ XY3$	+	+	
43.	$WX3\ WY3\ XY3$	+	+	
44.	$WX4\ WY3\ XY3$	+	+	
45.	$WX1\ WY4\ XY3$	+	+	
46.	$WX2\ WY4\ XY3$	+	+	
47.	$WX3\ WY4\ XY3$	+	+	
48.	$WX4\ WY4\ XY3$	+	+	
49.	$WX1\ WY1\ XY4$		+	+
50.	$WX2\ WY1\ XY4$	+	+	
51.	$WX3\ WY1\ XY4$	+	+	
52.	$WX4\ WY1\ XY4$	+	+	
53.	$WX1\ WY2\ XY4$		+	
54.	$WX2\ WY2\ XY4$	+	+	
55.	$WX3\ WY2\ XY4$	+	+	
56.	$WX4\ WY2\ XY4$	+	+	
57.	$WX1\ WY3\ XY4$		+	
58.	$WX2\ WY3\ XY4$	+	+	
59.	$WX3\ WY3\ XY4$	+	+	
60.	$WX4\ WY3\ XY4$	+	+	
61.	$WX1\ WY4\ XY4$		+	
62.	$WX2\ WY4\ XY4$	+	+	
63.	$WX3\ WY4\ XY4$	+	+	
64.	$WX4\ WY4\ XY4$	+	+	
65.	$WX1\ WY1\ XY5$		+	+
66.	$WX2\ WY1\ XY5$	+	+	
67.	$WX3\ WY1\ XY5$	+	+	
68.	$WX4\ WY1\ XY5$	+	+	
69.	$WX1\ WY2\ XY5$		+	
70.	$WX2\ WY2\ XY5$	+	+	
71.	$WX3\ WY2\ XY5$	+	+	
72.	$WX4\ WY2\ XY5$	+	+	
73.	$WX1\ WY3\ XY5$		+	
74.	$WX2\ WY3\ XY5$	+	+	

(*Continued on next page.*)

Table 4. (*Continued.*)

	Causal graph	$D(W, X)$	$D(X, Y)$	$I(W, Y \mid X)$
75.	$WX3\ WY3\ XY5$	+	+	
76.	$WX4\ WY3\ XY5$	+	+	
77.	$WX1\ WY4\ XY5$		+	
78.	$WX2\ WY4\ XY5$	+	+	
79.	$WX3\ WY4\ XY5$	+	+	
80.	$WX4\ WY4\ XY5$	+	+	
81.	$WX1\ WY1\ XY6$		+	+
82.	$WX2\ WY1\ XY6$	+	+	
83.	$WX3\ WY1\ XY6$	+	+	
84.	$WX4\ WY1\ XY6$	+	+	
85.	$WX1\ WY2\ XY6$	+	+	
86.	$WX2\ WY2\ XY6$	+	+	
87.	$WX3\ WY2\ XY6$	+	+	
88.	$WX4\ WY2\ XY6$	+	+	
89.	$WX1\ WY3\ XY6$	+	+	
90.	$WX2\ WY3\ XY6$	+	+	
91.	$WX3\ WY3\ XY6$	+	+	
92.	$WX4\ WY3\ XY6$	+	+	
93.	$WX1\ WY4\ XY6$	+	+	
94.	$WX2\ WY4\ XY6$	+	+	
95.	$WX3\ WY4\ XY6$	+	+	
96.	$WX4\ WY4\ XY6$	+	+	

Table 5. The three causal graphs in Table 4 that have the positive pattern + + +.

Causal graphs expressed as labels	Causal graphs expressed as networks
$WX2\ WY1\ XY2$	$W \rightarrow X \rightarrow Y$
$WX3\ WY1\ XY2$	H $W \quad X \rightarrow Y$
$WX4\ WY1\ XY2$	H $W \rightarrow X \rightarrow Y$

each of these graphs X causes Y. Thus, given Assumptions 1 through 7, a positive pattern of dependence and independence among three measured variables is sufficient to identify a causal relationship from among the 96 causal graphs modeled.

We note that since none of the relationships between X and Y in Table 5 are confounded by W, or by any hidden process, an estimate of $\Pr(Y \mid X)$ provides a direct estimate of the

distribution of Y given that we manipulate X to have some particular value (Spirtes et al., 1993).

The algorithm we describe in the next section is based on searching for positive patterns that exist among triplets of variables in $\boldsymbol{V}$. We emphasize, however, that it is possible for X to cause Y, and yet, a positive pattern will not hold; the causal graph $WX2\ WY2\ XY2$ is one such example, whereby W causes X, W causes Y, and X causes Y. Moreover, the algorithm we present in the next section does not guarantee returning all the possible causal constraints that could be identified from an observational database.

3.2. The LCD algorithm

This section specifies the LCD causal discovery algorithm and analyzes its computational time and space complexity.

In LCD, we assume the availability of a function called Independent(A, B, C) that uses complete database D (which we leave implicit) to determine whether A is independent of B, given C. If it is, then Independent returns *true*, otherwise, it returns *false*. If C is $\emptyset$, then Independent uses database D to determine if A is marginally independent of B. The function Independent might, for example, be based on classical statistical tests such as a chi-squared or a G^2 test of independence (Bishop et al., 1975). For such classical tests, we would need to specify the statistical threshold(s) needed to apply a given implementation of Independent. The appendix contains a Bayesian implementation of Independent, rather than an implementation that is based on a classical statistical test. The Bayesian implementation in the appendix is asymptotically correct in determining independence in the large sample limit. We also use a function Dependent(A, B, C), which is defined analogously to Independent. Its definition is given in the appendix as well.

The following pseudocode expresses the LCD algorithm. Curly brackets are used to enclose comments. We use "\" to denote the set difference operator.

```
procedure LCD(W, V, D);
  {Input: A variable W, which is assumed not to be caused by any other variable in the set
     V of discrete, measured variables in complete database D.}
  {Output: A printout of causal relationships that are discovered.}
  for X ∈ V \ {W} do
    if Dependent(W, X, ∅) then
          for Y ∈ V \ {W, X} do
                if Dependent(X, Y, ∅) then
                        if Independent(W, Y, X) then
                            write('the data support ', X, ' as a cause of ', Y);
                            {an estimate of the distribution Pr(Y | X)
                               could be printed here as well}
          end {for};
  end {for};
end {LCD}.
```

Every time LCD outputs that the data support X as a cause of Y, the algorithm also could output the distribution $\Pr(Y \mid X)$. Let y denote some value of Y and let x denote a value of X. In this context, the term $\Pr(Y = y \mid X = x)$ represents the probability that Y will take on the value Y given that we manipulate X to have the value x.

The analysis that follows assumes that the number of values of any variable in $\boldsymbol{V}$ is bounded from above by some constant. Since the analysis of Independent and Dependent are identical, we only discuss Independent. If there are n variables in $\boldsymbol{V}$ and m cases in D, then the time complexity of LCD is $O(n^2 f(m, n))$, where $f(m, n)$ is the time complexity of the function Independent. Typically, the time complexity of implementations of Independent would be $O(m)$, as is the case for the implementation in the appendix. If the complexity of Independent is $O(m)$, then the complexity of LCD is $O(n^2 m)$.

If there are at most q variables that test as dependent to W, where $q \geq 1$, then the time complexity of LCD is $O(nqf(m, n))$. If for some constant k, LCD only considers the k of n variables that have the strongest statistical dependency to W, then the time complexity of LCD is $O(nf(m, n))$. If the time complexity of Independent is $O(m)$, then the time complexity of LCD is $O(nm)$. By thus constraining LCD, we can make its time complexity proportional to the size of the database.

The space complexity of LCD is $O(mn)$, if we assume that LCD must store the database, which contains mn entries, and if we assume that the space complexity of Independent is $O(mn)$. More generally, the space complexity is $O(mn + g(m, n))$, where $g(m, n)$ is the space complexity of Independent.

Suppose there are a set of variables $\boldsymbol{U}$, each member of which we are willing to assume is not caused by any of the other variables in $\boldsymbol{V}$. We could call LCD for each W in $\boldsymbol{U}$. We note, but do not elaborate here, that when $|\boldsymbol{U}| > 1$, the use of additional data structures in LCD can lead to less redundant calls to Independent and Dependent, and thus to an improved efficiency of LCD.

3.3. Some limitations of LCD

This section discusses some types of causal relationships that LCD will miss, even when Assumptions 1 through 7 hold. We also illustrate that LCD returns pairwise causal relationships without unifying the causal relationships among those pairs.

There are 32 networks in Table 4 in which X is causing Y. The LCD algorithm is able (given that Assumptions 1 through 7 hold) to detect that X is causing Y in only 3 of these networks. The algorithm is unable to determine that X is causing Y in the other 29 networks, because the tested independence patterns associated with those networks are not unique for X causing Y. In particular, the 29 networks are numbered as 17, 21 through 32, and 65 through 80 in Table 4. Three representative examples are networks 22, 26, and 66, which are shown in figure 1.

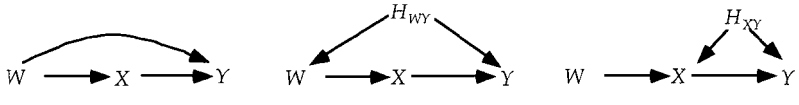

Figure 1. Networks 22 (left), 26 (middle), and 66 (right) from Table 4.

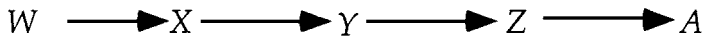

Figure 2. A hypothetical causal network.

We now consider one result of LCD returning separate pairwise causal relationships, rather than, for example, a single causal network that unifies all the causal pathways among the set of measured variables. Consider a causal process that is represented by the causal network in figure 2.

Given that Assumptions 1 through 7 hold, the LCD algorithm will not output the network in figure 2, but rather, it will output the following six pairwise causal relationships: $X \rightarrow Y$, $X \rightarrow Z, X \rightarrow A, Y \rightarrow Z, Y \rightarrow A, Z \rightarrow A$. Thus, true to its name, LCD focuses myopically on the discovery of local causal relationships, and therefore, it may produce a disjoint picture of the causal relationships. The local focus allows LCD to be simple and fast. Although not discussed here, post-processing of the LCD output could be used to produce a more unified summary of the causal relationships that are discovered.

Finally, we note that LCD need only condition on at most one variable when testing for independence; higher order tests of independence can be relatively less reliable.

4. The assumptions revisited

In this section, we revisit each of the seven assumptions on which LCD is based; we consider their plausibility and the implications of their failure. We also briefly suggest some possible extensions to LCD.

Assumption 1 (Database completeness). Often there is missing data in a database, that is, each variable is not measured for each case. One solution when considering variables W, X, and Y is to remove all cases from D in which any of these variables has a missing value. The problem with this approach is two-fold. First, we may end up with a very small database, and thus, have difficulty justifying Assumption 6 regarding valid statistical testing. Second, the data may not be missing randomly, in which case the distribution among the complete cases may not accurately reflect the distribution in the unselected population of interest; this effect can lead to a violation of Assumption 5 regarding no selection bias.

Another solution to the problem of missing data is to assign the value *missing* to a variable in a case for which that variable was not measured. This approach may, however, lead to an induced statistical dependency. In particular, conditioned on X having the explicit value *missing*, the value of W may provide some information about the value of Y, and thus, W and Y may test as being dependent; if this test result occurs, then LCD will miss uncovering that X causes Y.

A third solution to the problem is to *fill in* each missing value of each variable with some admissible value for the variable. There are numerous methods for assigning missing values (Little and Rubin, 1987). Hopefully, of course, the substituted values correspond closely to the actual, underlying values, but in general there is no guarantee that this will be the case.

Assumption 2 (Discrete variables). If variables are not discrete, then they can be discretized. If the granularity of the discretization is not fine enough (i.e., a variable is not given sufficient values), then induced statistical dependency may occur, for the same reasons discussed under Assumption 1 about modeling missing values. This situation may lead to missing the discovery of a causal relationship, but it should not itself cause the incorrect assertion of a causal relationship.

In principle, the LCD algorithm applies if there are continuous measured variables (or even a mixture of continuous and discrete variables), as long as there are functions Independent and Dependent that apply. For multivariate Gaussian Bayesian networks, such functions have been defined (Spirtes et al., 1993). Extending the Independent (Dependent) function to apply for a wide variety of distributions on continuous variables (or mixed continuous and discrete variables) is an open problem.

Assumption 3 (Bayesian network causal model). This assumption and the implied Markov condition are fundamental to our development and discussion of LCD thus far.

Although a detailed treatment is beyond the scope of this paper, let us consider briefly the possibility of a feedback loop between X and Y. Such a loop could be represented by a directed cycle, but such a representation would not be a Bayesian network. Consider extending the graphical representation presented thus far to include directed cycles, as is done by Richardson (1996) and by Pearl and Dechter (1996). Although not proved, we conjecture the following results: A positive pattern among W, X, and Y will exist only if X causes Y and Y does not cause X. Thus, the presence of feedback loops alone will not lead LCD to output incorrect causal assertions. Furthermore, in modeling the possibility of feedback loops, LCD will not miss making any causal assertions that it currently makes for causal processes that do not contain feedback loops.

Assumption 4 (Causal faithfulness condition). The most serious possible violation of the faithfulness condition in LCD would be the violation of the test Independent(W, Y, X). In Table 4, 71 of the 96 graphs (74%) have the pattern "$+$ $+$ ⟨blank⟩". Thus, if Independent(W, Y, X) erroneously returns true for any of these 71 graphs (while Dependent(W, X, $\emptyset$) and Dependent(X, Y, $\emptyset$) correctly return true), then LCD will erroneously conclude that X causes Y[3]. This is probably the place where LCD is the most vulnerable to error. One way to partially counter this vulnerability is to add the condition "Dependent(W, Y, $\emptyset$)" to LCD, which is a relationship that may fortuitously fail when Independent(W, Y, X) erroneously tests true. The total set of conditions tested in LCD would then be as follows:

Dependent(W, X, $\emptyset$) and Dependent(X, Y, $\emptyset$) and Dependent(W, Y, $\emptyset$) and Independent(W, Y, X)

An additional way to address the vulnerability is to use a more stringent statistical threshold for testing Independent(W, Y, X). Such an approach would decrease the likelihood of

LCD falsely reporting the existence of a causal relationship; it would, however, increase the likelihood of LCD missing the discovery of a valid causal relationship.

Assumption 5 (No selection bias). Suppose that an individual with only a fever (X) is likely to stay home and take aspirin. Similarly, a person with only abdominal pain (Y) is likely to stay home and take an over-the-counter medication for relief. Suppose, however, that an individual with both fever and abdominal pain is likely to be concerned about the possibility of a serious illness, and therefore, is prone to go to his or her local emergency room, where we have been collecting our data. In this situation, it is possible for X and Y to be dependent, due to selection bias, even though none of the relationships in Table 3 holds (Cooper, 1995). Such bias can persist, regardless of how large the sample size, and it may lead to LCD erroneously concluding that X causes Y.

Although a detailed treatment is beyond the scope of this paper, researchers have developed methods for detecting (at least in theory) the presence of selection bias (Spirtes et al., 1995). In short, given that Assumptions 1, 2, 3, 4, and 6 hold, then a modified version of LCD could avoid selection bias by only concluding that X causes Y if the following set of conditions holds:

$$\begin{aligned}&\text{Dependent}(W1,\, X, \varnothing) \text{ and Dependent}(W2,\, X, \varnothing) \text{ and Independent}(W1,\, W2, \varnothing)\\ &\quad\text{and Dependent}(X,\, Y, \varnothing) \text{ and Independent}(W1,\, Y,\, X). \qquad (1)\end{aligned}$$

In essence, if the dependency between X and Y is due at least in part to selection bias, then $W1$ and $W2$ would be expected to be dependent (although in reality they may not be dependent, due to a violation of one or more of the other assumptions).

Assumption 6 (Valid statistical testing). The smaller the number of cases in D, the more skeptical we should be of whether Assumption 6 holds. Even for a large database, however, it is not clear which value to use as a statistical threshold for a classical test of independence, such as the chi-squared test. The Bayesian version of Independent (Dependent), which is described in the appendix, returns the correct answer in the large sample limit. Thus, the issue of which particular probability threshold to use is of relatively less concern, but nonetheless, it remains a relevant issue.

Assumption 7 (An uncaused entity). If W is caused by X, then it is possible to conclude that X causes Y, when in fact Y causes X. Thus, it is important that we choose W carefully.

In expression 1 above, we do not need to assume that $W1$ and $W2$ are two variables that are not caused by any other variables in $\boldsymbol{V}$ (Spirtes et al., 1993). That is, expression 1 is sufficient to determine that X causes Y (given that Assumptions 1, 2, 3, 4, and 6 hold), even in the absence of assuming that $W1$ and $W2$ have no measured causes. Thus, it is possible to discover causal relationships in the absence of any known or assumed causal relationships among a set of measured variables. At times, just the observational measurements themselves can reveal causation.

We now discuss the testability of Assumptions 1 through 7. The validity of Assumptions 1 (database completeness) and 2 (discrete variables) are readily apparent. Assumptions 3 (Bayesian network causal model), 4 (causal faithfulness condition), and 6 (valid statistical testing) are not readily testable. The conditions in expression 1 provide a test of whether Assumption 5 (no selection bias) is valid, subject to Assumptions 1, 2, 3, 4, and 6 being valid. In any given application, however, those conditions may not be met. Assumption 7 (an uncaused event) is potentially testable, but requires knowledge outside of database D and outside of Assumptions 1 through 6. For example, we may have knowledge about time precedence that allows us to conclude the presence of an uncaused event.

5. Other algorithms for causal discovery

In this section, we discuss some selected, representative, prior research on constraint-based causal discovery and Bayesian causal discovery, and we briefly relate this work to LCD.

5.1. The PC and FCI algorithms

LCD can be viewed as a specialization of the PC and FCI constraint-based causal discovery algorithms that uses background knowledge about an uncaused variable (W). The PC and FCI algorithms are described in detail in (Spirtes et al., 1993) and are available commercially (Scheines et al., 1995). While these two algorithms certainly are more difficult to implement than LCD, in an absolute sense they are not especially difficult to implement.

PC assumes no hidden variables, while FCI allows hidden variables. Both algorithms search for causal constraints among all the variables in $\boldsymbol{V}$, rather than restrict search to triplets of variables, as does LCD. Consequently, they are able to find causal relationships that LCD misses. Both PC and FCI have a richer language than LCD for constraining the causal relationships that exist among a set of variables. The FCI algorithm, for example, has a constraint language that includes the following predicates: (1) X causes Y, (2) X is not caused by Y, (3) X and Y have a common hidden variable causally influencing each of them, and (4) X is either a cause of Y or a cause of Z. On the other hand, LCD only makes assertions of type 1. Moreover, PC and FCI can output valid assertions of type 1 that LCD misses.

PC and FCI also provide a relatively unified model of causality. So, for example, unified causal networks similar to that in figure 2 could be output by PC and FCI, in contrast to the set of pairwise causal relationships that would be output by LCD, as described in Section 3.3.

The price PC and FCI pay for their generality is that in the worst case their search time can be exponential in the number of variables in $\boldsymbol{V}$, unlike LCD, which is polynomial time. Also, both PC and FCI may test for independence relationships based on a large number of conditioning variables; such tests tend to be less reliable than the low order independence tests used in LCD.

PC and FCI use Assumption 3 (Bayesian network causal model), Assumption 4 (faithfulness condition), and Assumption 6 (valid statistical testing). A practical application of the algorithms typically involves deleting cases in which any variable has a missing value, but

as discussed in Section 4, this approach can lead to selection bias. In theory, however, the PC and FCI methods can detect such induced selection bias (Spirtes et al., 1995), and thus, avoid reporting incorrect causal constraints. A more serious problem with case deletion is that the sample size of many real-world databases may become very small, thus jeopardizing the validity of Assumption 6.

5.2. Bayesian causal discovery

Bayesian methods have been developed for discovering causal relationships from observational data (Cooper and Herskovits, 1992; Heckerman, 1996; Heckerman et al., 1995). These methods differ in several ways from constraint-based methods. First, the methods take a user-specified prior probability over Bayesian network structures and parameters. If the user has little prior information, or it is not feasible to specify this information, then non-informative priors can be used. The methods then can return a posterior probability over one or more causal graphs and/or over one or more causal arcs. No statistical testing thresholds need to be specified; this property is welcome, since the thresholds applied in constraint-based methods are chosen somewhat arbitrarily.

When Assumptions 1 through 6 hold, and when there are no hidden variables, in the large sample limit the Bayesian methods and PC will identify the same set of causal relationships among the measured variables (Bouckaert, 1995). If there are hidden variables, however, the Bayesian methods can make distinctions that PC and FCI cannot make. For example, the Bayesian methods sometimes can determine the likely number of values for a hidden variable.

One primary problem with Bayesian methods is computational tractability, because an exact computation with current algorithms requires summing over a number of causal graphs that is exponential in the number of graph variables. In simulation experiments, however, the application of Bayesian methods with heuristic search techniques has been effective in recovering causal structure on measured variables (Aliferis and Cooper, 1994; Cooper and Herskovits, 1992; Heckerman et al., 1995). When there are hidden variables, exact computation with current Bayesian methods often is intractable, even when the causal graphs contain only a few variables. The use of sampling and approximation methods, however, recently has shown promise (Chickering and Heckerman, 1996). In summary, even though exact application of Bayesian methods often is intractable, approximate solutions may be acceptable.

Another challenge of applying Bayesian methods for causal discovery is the assessment of informative priors on possible causal structures and on parameters for those structures. On the one hand, the ability to represent such prior information is a great strength of the Bayesian approach. With it, we can potentially express prior causal knowledge that comes from other sources, such as experiments, observational experience, common sense, and physical constraints. While good progress has been made in facilitating the expression of priors on Bayesian network structures and parameters (Heckerman et al., 1995), assessing such prior probabilities (particularly when there is a large set of variables) can still be difficult and sometimes infeasible. Thus, currently, it is common to specify some form of a non-informative prior on the causal structures (e.g., a uniform prior over all possible

structures) and on the parameters of those structures. Non-informative priors typically require that the user specify only a few parameters; still, it sometimes is not obvious what these few parameters should be. In that case, performing a sensitivity analysis over the parameters may be a good idea.

Although significant research challenges remain in making the Bayesian approach feasible from a computational and an assessment standpoint, it is the author's opinion that the strengths of the approach, as summarized in this section, will lead ultimately to it (or a combination of it and constraint-based approaches) being the most commonly applied class of causal discovery methods. This paper presents the constraint-based approach in more detail, because it is a viable method of causal discovery that—in its most basic form—is particularly easy to convey and program.

We conclude this section by describing a straightforward Bayesian version of LCD. For a given triplet of variables, we specify a prior over all 96 causal graphs in Table 4, and for each graph (i.e., Bayesian network structure), we specify a prior over the probability distributions in that graph[4]. Alternatively, a much quicker approach would be to use a non-informative prior (Cooper and Herskovits, 1992; Heckerman et al., 1995). The probability $\Pr(S_i, D)$ would be computed for each of the 96 causal graphs; call $\Pr(S_i, D)$ the score for graph S_i (see the Appendix for one method to compute $\Pr(S_i, D)$). For those graphs containing hidden variables, it likely would be necessary to apply an approximation method to compute a score in a feasible amount of time (Chickering and Heckerman, 1996). Let t be the sum of scores over all graphs that contain an arc from X to Y. Let u be the sum of scores over all 96 graphs. Then $\Pr(X \rightarrow Y \mid D, \xi) = t/u$, where ξ denotes all the priors and assumptions applied.

The problem with a Bayesian version of LCD is that it does not consider the relationships of all the variables in $\boldsymbol{V}$ at once. As with LCD, a focus on searching over triplets of variables will gain computational efficiency, but lose the ability to identify some causal relationships. Moreover, the Bayesian posterior probabilities for a triplet of variables will only be strictly correct if we focus on a single triplet, which is unlikely to happen in a data mining application. Corrections are possible, but they are beyond the scope of this paper.

6. Discussion

We have presented a simple, efficient algorithm called LCD for causal discovery from observational data. We listed seven assumptions on which LCD is based. While we can weigh factors for and against each of these assumptions, it is difficult to imagine that a theoretical argument will determine their effect in combination. Ultimately, the most interesting question is the utility, rather than the validity, of their combined application. We believe that an assessment of real-world utility must rest on real-world, empirical results. If causal discovery programs that are based on these assumptions are applied to multiple databases and are helpful in finding causal relationships, then we will have a useful set of working assumptions. This is an important empirical issue that needs to be addressed much more extensively. Hopefully, this paper will encourage readers to apply causal discovery methods to their data (Almond, 1997; Heckerman, 1996; Scheines et al., 1995), and thereby help to determine the real-world utility of the methods.

Appendix

This appendix describes one possible implementation of the Independent and Dependent functions used in LCD. Both functions use a previously developed Bayesian scoring metric. In the pseudocode that follows, square brackets are used to delineate a Bayesian network structure. The function $\Pr(S, D)$ is described after the listing of the pseudocode.

function Independent(X_1, X_2, X_3): boolean;
{Input: A database D, which is global. Variables X_1, X_2, and X_3, which represent nodes.}
{Output: Return the value true if X_1 is (to some tolerance) independent of X_2 given X_3, otherwise return the value false. Variable X_3 may be nil or it may represent a single node.}
  $t := 0.9$; {This is a user-specified probability threshold. Arbitrarily it is set to 0.9 here.}
  if $X3 = \varnothing$ **then**
    $S := [X_1 <\text{no\_arc}> X_2]$; {The structure with no arc from X_1 to X_2.}
    $a := \Pr(S, D)$;
    $S := [X_1 \rightarrow X_2]$; {The structure with an arc from X_1 to X_2.}
    $b := \Pr(S, D)$;
    if $a/(a + b) > t$ **then** Independent := true **else** Independent := false;
  else
    $S := [X_1 <\text{no\_arc}> X_3, X_3 <\text{no\_arc}> X_2, X_1 <\text{no\_arc}> X_2]$;
    $a := \Pr(S, D)$;
    $S := [X_1 \rightarrow X_3, X_3 <\text{no\_arc}> X_2, X_1 <\text{no\_arc}> X_2]$;
    $b := \Pr(S, D)$;
    $S := [X_1 <\text{no\_arc}> X_3, X_3 \rightarrow X_2, X_1 <\text{no\_arc}> X_2]$;
    $c := \Pr(S, D)$;
    $S := [X_1 \rightarrow X_3, X_3 \rightarrow X_2, X_1 <\text{no\_arc}> X_2]$;
    $d := \Pr(S, D)$;
    $S := [X_1 <\text{no\_arc}> X_3, X_3 <\text{no\_arc}> X_2, X_1 \rightarrow X_2]$;
    $e := \Pr(S, D)$;
    $S := [X_1 \rightarrow X_3, X_3 <\text{no\_arc}> X_2, X_1 \rightarrow X_2]$;
    $f := \Pr(S, D)$;
    $S := [X_1 <\text{no\_arc}> X_3, X_3 \rightarrow X_2, X_1 \rightarrow X_2]$;
    $g := \Pr(S, D)$;
    $S := [X_1 \rightarrow X_3, X_3 \rightarrow X_2, X_1 \rightarrow X_2]$;
    $h := \Pr(S, D)$;
    if $(a + b + c + d)/(a + b + c + d + e + f + g + h) > t$
    then Independent := true
    else Independent := false;
end {Independent}.

The function Dependent is very similar to Independent, and the pseudocode that follows focuses on their differences.

function Dependent(X_1, X_2, X_3): boolean;
  $t := 0.9$; {This is a user-specified probability threshold. Arbitrarily it is set to 0.9 here.}
  if $X_3 = \varnothing$ **then**
    ...
    if $b/(a+b) > t$ **then** Dependent := true **else** Dependent := false;
  else
    ...
    if $(e+f+g+h)/(a+b+c+d+e+f+g+h) > t$ **then** Dependent := true
    else Dependent := false;
end {Dependent}.

One Bayesian metric for computing $\Pr(S, D)$, which is derived in Cooper and Herskovits (Herskovits, 1991), is as follows. Let $\boldsymbol{Z}$ be a set of n discrete variables, where a variable X_i in $\boldsymbol{Z}$ has r_i possible value assignments: $(v_{i1}, \ldots, v_{ir_i})$. Let D be a database of m cases, where each case contains a value assignment for each variable in $\boldsymbol{Z}$. Let S denote a Bayesian network structure containing just the variables in $\boldsymbol{Z}$. Each variable X_i in S has a set of parents, which we represent with a list of variables π_i. Let w_{ij} denote the jth unique instantiation of π_i relative to D. Suppose there are q_i such unique instantiations of π_i. Define N_{ijk} to be the number of cases in D in which variable X_i has the value v_{ik} and π_i is instantiated as w_{ij}. Let $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Given the assumptions made in (Cooper and Herskovits, 1992), it follows that

$$\Pr(S, D) = \Pr(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!.$$

For the purpose to which $\Pr(S, D)$ is applied in Independent and Dependent, the term $\Pr(S)$ in the above equation may be simply set to 1. Herskovits (1991) contains an analysis of the convergence properties of the above scoring metric. By using the results of this analysis, it can be shown that in the large sample limit the above implementation of Independent(X_1, X_2, X_3) returns *true* if and only if X_1 is independent of X_2 given X_3. A similar analysis holds for Dependent.

Acknowledgments

I thank Constantin Aliferis, John Aronis, Clark Glymour, Stefano Monti, Peter Spirtes, Cleat Szczepaniak, and the anonymous reviewers for their excellent comments on earlier versions of this paper. This research was supported in part by grant BES-9315428 from the National Science Foundation and by grant LM05291-02 from the National Library of Medicine.

Notes

1. Note that although the representation we use is called a *Bayesian* network, the method that we use to learn such networks will be constraint-based, rather than Bayesian. In Section 5.2 we briefly discuss Bayesian methods for learning Bayesian networks.

2. For an interactive tutorial on *d*-separation, see http://www.andrew.cmu.edu/user/wimberly/dsep/dSep.html.
3. Such a violation typically would occur when a distribution is "close" to being unfaithful and/or the data sample is small. Thus, in practice, we are really discussing here an interplay of Assumptions 4 and 6.
4. As mentioned in Section 3.1, there actually are more than 96 possible graphs, when considering three measured variables plus hidden variables. In general, then, we would need to consider all such graphs for which we have a prior probability that is greater than zero.

References

Aliferis, C.F. and Cooper, G.F. 1994. An evaluation of an algorithm for inductive learning of Bayesian belief networks using simulated data sets. Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 8–14.

Almond, R.G., 1997. Web page on Software for Learning Belief Networks from Data, http://bayes.stat.washington.edu/almond/belfit.html#BNG.

Bishop, Y., Fienberg, S., and Holland, P. 1975. Discrete Multivariate Analysis. Cambridge, MA: MIT Press.

Bouckaert, R. 1995. Bayesian belief networks: From construction to inference, Doctoral dissertation, University of Utrecht, Utrecht, Netherlands.

Castillo, E., Gutierrez, J.M., and Hadi, A.S. 1997. Expert Systems and Probabilistic Network Models. New York: Spring-Verlag.

Chickering, D.M. and Heckerman, D. 1996. Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 158–168.

Cooper, G.F. 1995. Causal discovery from data in the presence of selection bias. Proceedings of the Workshop on Artificial Intelligence and Statistics, pp. 140–150.

Cooper, G.F. and Herskovits, E. 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9:309–347.

Geiger, D., Verma, T., and Pearl, J. 1990. Identifying independence in Bayesian networks. Networks 20:507-–534.

Heckerman, D. 1996. A tutorial on learning with Bayesian networks, Microsoft Research Report MSR-TR-95-06 (available at http://www.research.microsoft.com/ research/dtg/heckerma/heckerma.html).

Heckerman, D., Geiger, D., and Chickering, D. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 20:197–243.

Herskovits, E.H. 1991. Computer-based probabilistic-network construction, Doctoral dissertation, Medical Information Sciences, Stanford University.

Jensen, F.V. 1996. An Introduction to Bayesian Networks. New York: Springer-Verlag.

Little, R.J.A. and Rubin, D.B. 1987. Statistical Analysis with Missing Data. New York: John Wiley & Sons.

Meek, C. 1995. Strong completeness and faithfulness in Bayesian networks. Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 411–418.

Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems. San Mateo, CA: Morgan Kaufmann.

Pearl, J. 1994. Causal diagrams for empirical research, Report R-218-L, Computer Science Department, University of California at Los Angeles.

Pearl, J. and Verma, T.S. 1991. A theory of inferred causality. Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning, pp. 441–452.

Pearl, J. and Dechter, R. 1996. Identifying independencies in causal graphs with feedback. Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 420–426.

Richardson, T. 1996. A discovery algorithm for directed causal graphs. Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 454–461.

Scheines, R., Spirtes, P., Glymour, C., and Meek, C. 1995. Tetrad II: Tools for Causal Modeling (with software). Mahwah, New Jersey: Lawrence Erlbaum.

Spirtes, P., Glymour, C., and Scheines, R. 1993. Causation, Prediction, and Search. New York: Springer-Verlag. (This book is out of print, but it is available in its entirety in Adobe Acrobat format at http://hss.cmu.edu/html/departments/philosophy/TETRAD.BOOK/book.html).

Spirtes, P., Meek, C., and Richardson, T. 1995. Causal inference in the presence of latent variables and selection bias. Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 499–506.

Gregory F. Cooper, M.D., Ph.D. is an Associate Professor of Medicine at the University of Pittsburgh with a joint academic appointment in the Intelligent Systems Program. He also is a member of the Center for Biomedical Informatics. His primary research interests involve the application of Bayesian statistics, decision theory and artificial intelligence to a variety of research problems, including causal discovery from observational data, computer-aided clinical outcome prediction, computer-assisted development of clinical care guidelines, and machine learning of expert systems.