Journal of Economic Methodology 12:1, 1–33 March 2005

Routledge
Taylor & Francis Group

Graphical models, causal inference, and econometric models

Peter Spirtes

Abstract A graphical model is a graph that represents a set of conditional independence relations among the vertices (random variables). The graph is often given a causal interpretation as well. I describe how graphical causal models can be used in an algorithm for constructing partial information about causal graphs from observational data that is reliable in the large sample limit, even when some of the variables in the causal graph are unmeasured. I also describe an algorithm for estimating from observational data (in some cases) the total effect of a given variable on a second variable, and theoretical insights into fundamental limitations on the possibility of certain causal inferences by any algorithm whatsoever, and regardless of sample size.

Keywords: graphical models, causal inference, model search, model testing

1 INTRODUCTION

A graphical model consists of a graph with vertices that are random variables, and an associated set of joint probability distributions over the random variables, all of which share a set of conditional independence relations. The graph is often given a causal interpretation as well, in which case it is a graphical causal model. Linear structural equation models with associated path diagrams are examples of graphical causal models. By exploiting the relationship between graphs and conditional independence relations on the one hand, and graphs and causal relations on the other hand, many properties of path diagrams can be generalized to a wide variety of families of distributions, and assumptions about linearity can be relaxed.

Although graphical causal modeling has historical ties to causal modeling in econometrics and other social sciences, there have been recent developments by statisticians, computer scientists, and philosophers that have been relatively isolated from the econometric tradition. In this paper I will describe a number of recent developments in graphical causal modeling, and their relevance to econometrics and other social sciences.

The use of graphs to represent both causal relations and sets of conditional independence relations (which will henceforth be referred to

Journal of Economic Methodology ISSN 1350-178X print/ISSN 1469-9427 online © 2005 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/1350178042000330887

as the graphical approach to causal inference) is relevant to econometric and other social science methodology for three kinds of reasons. First, the graphical approach to causal inference has led to a more explicit formulation of the assumptions implicit in some social science methodology (see, for example, sections 2.3 and 4). This in turn enables one to examine when the assumptions (and associated methods) are reasonable and when they are not.

Second, the graphical approach to causal inference has led to the discovery of a number of useful algorithms. These include, among others, algorithms for searching for partial information about causal graphs that are reliable in the large sample limit (the sense of 'reliability' and the assumptions under which the algorithms are reliable are described in section 4), and an algorithm for estimating the total effect of a given variable on a second variable in some cases, given partial information about causal graphs, even when some of the variables in the graph are unmeasured (sections 4.4 and 5).

Third, the graphical approach to causal inference has led to theoretical insights into fundamental limitations on the possibility of certain causal inferences by any algorithm whatsoever, and regardless of sample size. One fundamental limitation on causal inference from observational data is the underdetermination of causal models by probability distributions, i.e. more than one causal model is compatible with a given probability distribution. Using graphical models, the extent of this underdetermination can be precisely characterized, under assumptions relating causal models to probability distributions (section 4).

Section 2 describes in more detail a simple kind of graphical model, a linear structural equation model (LSEM) that illustrates many of the basic concepts; section 3 describes the main obstacles to reliable causal modeling and a standard of success that can be applied to judge whether an algorithm for causal inference is 'reliable' or not; section 4 describes assumptions and algorithms that provide reliable causal inference in LSEMs; section 5 sketches how to extend the same basic ideas to latent variables models; section 6 mentions some extensions to cyclic models and VAR models; section 7 describes properties of the search algorithm at finite sample sizes; and section 8 is the conclusion.

2 LSEM MODELING

LSEMs are a special case of the much broader class of graphical models. Answers to questions about searching for and selecting LSEMs will shed light on the problem of searching for and selecting other more realistic models.

2.1 The statistical interpretation of LSEMs

In an LSEM the random variables are divided into two disjoint sets, the substantive variables (typically the variables of interest) and the error variables (summarizing all other variables that have a causal influence on the substantive variables). Corresponding to each substantive random variable V is a unique error term ε_V. An LSEM contains a set of linear equations in which each substantive random variable V is written as a linear function of other substantive random variables together with ε_V, a correlation matrix among the error terms, and the means of the error terms. Initially, it will be assumed that the error variables are multivariate normal. However, many of the results that are proved are about partial correlations, which do not depend upon the distribution of the error terms, but depend only upon the linearity of the equations and the correlations among the error terms.

The only LSEMs that will be considered are those that have coefficients for which there is a reduced form (i.e. all substantive variables can be written as functions of error terms alone), all variances and conditional variances among the substantive variables are finite and positive, and all conditional correlations among the substantive variables are well defined (e.g. not infinite).

The path diagram of an LSEM with uncorrelated errors is written with the conventions that it contains an edge $A \rightarrow B$ if and only if the coefficient for A in the structural equation for B is non-zero, and there is a double-headed arrow between two error terms ε_A and ε_B if and only if the correlation between ε_A and ε_B is non-zero. ('Path diagram' and 'graph' will be used interchangeably in what follows. What makes these equations 'structural' is described in section 2.2.) An error term that is not correlated with any other error term is not included in the path diagram.

The following model is an example of an LSEM with free parameters. The path diagram for the LSEM is G_1 of Figure 1. A *directed graph* consists of a set of vertices and a set of directed edges, where each edge is an ordered pair of vertices. In G_1, the vertices are $\{A,B,C,D,E\}$, and the edges are $\{B \rightarrow A, B \rightarrow C, D \rightarrow C, C \rightarrow E\}$. In G_1, B is a *parent* of A, A is a *child* of B, and A and B are *adjacent* because there is an edge $A \rightarrow B$. A *path* in a directed graph is a sequence of adjacent edges (i.e. edges that share a single common endpoint). A *directed path* in a directed graph is a sequence of adjacent edges all pointing in the same direction. For example, in G_1, $B \rightarrow C \rightarrow E$ is a directed

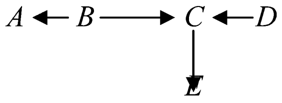

Figure 1 G_1 of Model 1

path from B to E. In contrast, $B \rightarrow C \leftarrow D$ is a path, but not a directed path in G_1 because the two edges do not point in the same direction; in addition, C is a *collider on the path* because both edges on the path are directed into C. A triple of vertices $<B,C,D>$ is a *collider* if there are edges $B \rightarrow C \leftarrow D$ in G_1; $<B,C,D>$ is an *unshielded* collider if in addition there is no edge between B and D. E is a *descendant* of B (and B is an *ancestor* of E) because there is a directed path from B to E; in addition, by convention, each vertex is a descendant (and ancestor) of itself. A directed graph is *acyclic* when there is no directed path from any vertex to itself: in that case the graph is a directed acyclic graph, or DAG for short.

The structural equations are: $A := \theta_{BA}B + \varepsilon_A$; $B := \varepsilon_B$; $C := \theta_{BC}B + \theta_{DC}D + \varepsilon_C$; $D := \varepsilon_D$; and $E := \theta_{CE}C + \varepsilon_E$. θ_{BA}, θ_{BC}, θ_{DC}, and θ_{CE} are free parameters of the model, which take on any real values except zero (which is excluded in order to ensure that a model and its submodels are disjoint). For reasons explained in section 2.2, following the notation of Lauritzen (2001), an assignment operator ':=' rather than an equals sign is used, to emphasize that the equation is a structural equation. The other free parameters are the variances and means of the error terms ε_A, ε_B, ε_C, ε_D, and ε_E, which are denoted by σ_A, σ_B, σ_C, σ_D, σ_E, and μ_A, μ_B, μ_C, μ_D, and μ_E respectively. The set of free parameters is denoted as $\mathbf{\Theta_1} = <\theta_{BA}, \theta_{BC}, \theta_{DC}, \theta_{CE}, \sigma_A, \sigma_B, \sigma_C, \sigma_D, \sigma_E, \mu_A, \mu_B, \mu_C, \mu_D, \mu_E>$, and the model with free parameters as $<G_1, \mathbf{\Theta_1}>$.

If specific values are assigned to the free parameters, e.g. $\Theta_1 = <\theta_{BA}=2, \theta_{BC}=0.6, \theta_{DC}=-0.4, \theta_{CE}=1.3, \sigma_A=1, \sigma_B=1, \sigma_C=1, \sigma_D=1, \sigma_E=1, \mu_A=0, \mu_B=0, \mu_C=0, \mu_D=0, \mu_E=0>$, then the resulting parameterized model is $<G_1, \Theta_1>$. The structural equations are: $A := 2B + \varepsilon_A$; $B := \varepsilon_B$; $C := 0.6B - 0.4D + \varepsilon_C$; $D := \varepsilon_D$; and $E := 1.3C + \varepsilon_E$.

The covariance matrix over the error terms (the non-diagonal terms are zero because there are no double-headed arrows in G_1), together with the linear coefficients, determine a unique covariance matrix over the substantive variables A, B, C, D, and E. For a particular pair $<G_1, \Theta_1>$ the corresponding distribution is denoted as $f(<G_1, \Theta_1>)$. The range of some parameters must be restricted in order for the parameters to specify a probability distribution; e.g. the standard deviations cannot be negative. In addition, in order to make submodels disjoint from supermodels, the linear coefficient free parameters are restricted to non-zero values. Any parameter value that falls within the restricted range of the parameters will be referred to as a 'legal' parameter value. The set of all distributions corresponding to legal values of the parameters is denoted as $\mathbf{P}(<G_1, \Theta_1>)$.

In *all* of the distributions in $\mathbf{P}(<G_1, \mathbf{\Theta_1}>)$ some conditional independence relations hold (i.e. they are entailed to hold for all legal values of the parameters). The set of conditional independence relations that holds in every distribution in $\mathbf{P}(<G_1, \mathbf{\Theta_1}>)$ is denoted by $\mathbf{I}(<G_1, \mathbf{\Theta_1}>)$. In a multivariate normal distribution, the partial correlation $\rho(X,Y|\mathbf{Z})$ is zero if and only if X and Y are independent conditional on $\mathbf{Z}$. So for multivariate normal distributions, the conditional independence relations in $\mathbf{I}(<G_1, \boldsymbol{\Theta}_\mathbf{1}>)$ can be specified by listing a set of zero partial correlations among the variables.[1]

There are many different ways of parameterizing a graph such as G_1. Because G_1 is a DAG (i.e. it contains directed edges, but no directed cycles and no bidirected edges), the LSEM parameterization has the property that it entails that each variable in G_1 is independent of the variables that are neither its descendants nor its parents, conditional on its parents. Any probability distribution that satisfies this property for G_1 is said to satisfy the **local directed Markov property** for G_1.

In the multivariate normal case, the following partial correlations are entailed to equal zero by the local directed Markov property for G_1: $\rho(A,C|\{B\})$, $\rho(A,D|\{B\})$, $\rho(A,E|\{B\})$, $\rho(B,D)$, $\rho(C,A|\{B,D\})$, $\rho(D,A)$, $\rho(D,B)$, $\rho(E,A|\{C\})$, $\rho(E,B|\{C\})$, and $\rho(E,D|\{C\})$. The list contains some redundancies [e.g. $\rho(B,D)$ and $\rho(D,B)$] because it contains all applications of the local directed Markov property to each variable in the graph, whether they are redundant or not.

The conditional independence relations entailed by satisfying the local directed Markov property in turn entail all of the other conditional independence relations in $\mathbf{I}(<G_1, \boldsymbol{\Theta}_\mathbf{1}>)$. For any distribution that satisfies the local directed Markov property for G_1, all of the conditional independence relations in $\mathbf{I}(<G_1, \boldsymbol{\Theta}_\mathbf{1}>)$ hold. Since these independence relations don't depend upon the particular parameterization but only on the graphical structure and the local directed Markov property, they will henceforth be denoted by $\mathbf{I}(G_1)$.

There is an (unfortunately unintuitive) graphical relationship among sets of variables in a DAG G named 'd-separation' that determines which conditional independence relations belong to $\mathbf{I}(<G_1, \boldsymbol{\Theta}_\mathbf{1}>)$ (i.e. are entailed by satisfying the local directed Markov property). Following Pearl (1988), in $\boxed{1}$ a DAG G, for disjoint variable sets $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, $\mathbf{X}$ and $\mathbf{Y}$ are **d-separated** conditional on $\mathbf{Z}$ in G if and only if there exists no path U between an $X \in \mathbf{X}$ and a $Y \in \mathbf{Y}$ such that (i) every collider on U has a descendent in $\mathbf{Z}$; and (ii) no other vertex on U is in $\mathbf{Z}$. A DAG G entails that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ (in the multivariate normal case $\rho(X,Y|\mathbf{Z})=0$ for all $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$) if and only if $\mathbf{X}$ is d-separated from $\mathbf{Y}$ conditional on $\mathbf{Z}$ in G.

For multivariate normal distributions, the set of d-separation relations between pairs of variables in G_1 corresponds to the set of partial correlations entailed to be zero by the local directed Markov property. For G_1, the complete set of partial correlations entailed to be zero is: $\{\rho(A,C|\{B\})$, $\rho(A,C|\{B,D\})$ $\rho(A,C|\{B,E\})$, $\rho(A,C|\{B,D,E\})$, $\rho(A,D)$, $\rho(A,D|\{B\})$, $\rho(A,D|\{B,C\})$, $\rho(A,D|\{B,E\})$, $\rho(A,D|\{B,C,E\})$, $\rho(A,E|\{C\})$, $\rho(A,E|\{B\})$, $\rho(A,E|\{B,C\})$, $\rho(A,E|\{B,D\}$, $\rho(A,E|\{C,D\})$, $\rho(A,E|\{B,C,D\})$, $\rho(B,D)$, $\rho(B,D|\{A\})$,

$\rho(B,E|\{C\})$, $\rho(B,E|\{C,D\})$, $\rho(B,E|\{A,C\})$, $\rho(B,E|\{A,C,D\})$, $\rho(D,E|\{C\})$, $\rho(D,E|\{A,C\})$, $\rho(D,E|\{B,C\})$, $\rho(D,E|\{A,B,C\})\}$.

For multivariate Normal distributions, every probability distribution that satisfies the set of conditional independence relations in $\mathbf{I}(<G_1, \mathbf{\Theta_1}>)$ is also a member of $\mathbf{P}(<G_1, \mathbf{\Theta_1}>)$. However, for other families of distributions, it is possible that there are distributions that satisfy the conditional independence relations in $\mathbf{I}(<G_1, \mathbf{\Theta_\alpha}>)$, but are not in $\mathbf{P}(<G_1, \mathbf{\Theta_\alpha}>)$ (i.e. the parameterization imposes constraints that are not conditional independence constraints). See Lauritzen *et al.* (1990), Pearl (2000), or Spirtes, Glymour and Scheines (2000) for details.

2.2 The causal interpretation of LSEMs

The conditional distribution $f(B|A=a)$ represents the probability distribution of B in a subpopulation in which $A=a$, or when A has been *observed* to have value a, but the causal system has not been interfered with. The density of B when A is *manipulated* to have value a represents the probability distribution of B in a hypothetical population in which an manipulation has been performed on each member of the population to *set* the value of A to a. Conditioning is typically the appropriate operation when attempting to diagnose the value of a hidden variable, or to predict the future. Manipulating is typically the appropriate operation when calculating the effect of adopting some policy.

For example, when attempting to determine the probability that someone who does not have nicotine-stained fingers has lung cancer it would be appropriate to use *f*(*Lung Cancer*=*yes*|*Nicotine-stained fingers*=*no*). In guessing whether *Lung Cancer*=*yes* from the value of *Nicotine-stained fingers*, it does not matter whether *Nicotine-stained fingers* is an effect of *Lung Cancer*, a cause of *Lung Cancer*, or (as is actually the case) *Lung Cancer* and *Nicotine-stained fingers* have a common cause (*Smoking*). On the other hand, if one is considering a plan to reduce the incidence of *Lung Cancer* by encouraging people to remove the nicotine stains from their fingers, then the relevant question is whether manipulating the system to wash nicotine stains off of fingers is going to reduce the incidence of lung cancer; the quantity that represents this is the distribution of *Lung Cancer*=*yes* when the population is manipulated to have *Nicotine-stained fingers*=*no*. The manipulated distribution does depend upon whether *Nicotine-stained fingers* is an effect of *Lung Cancer*, a cause of *Lung Cancer*, or *Lung Cancer* and *Nicotine-stained fingers* have a common cause. In this example, it is intuitively clear that *f*(*Lung Cancer*=*yes*|*Nicotine-stained fingers*=*no*) is not the same as the distribution of *Lung Cancer*=*yes* when the population is manipulated to have *Nicotine-stained fingers*=*no* (because the latter group contains both smokers and non-smokers.).

In the rest of this section, notation to describe manipulated distributions will be introduced, as well as an explanation of how manipulations can be represented in LSEM. No definition of ‘direct cause’ is provided here, but an axiomatic approach is adopted that makes assumptions about the relationships between direct causation and probability distributions, and the relationship between direct causes and the effects of manipulating variables in a causal system.

Under the causal interpretation, the equations in an LSEM are ‘structural’ because the variables on the right hand side of the equation are direct causes of the variables on the left hand side of the equation, and the equations can be used to calculate the effects of manipulating the variables in the system. As a result, there is an edge from X to Y in the corresponding path diagram just when X is a direct cause of Y. Intuitively, one simple kind of manipulation of a variable A is a randomized experiment on A which replaces the naturally occurring distribution of A with a new distribution that is imposed upon it. More complex manipulations allow randomized experiments to be performed upon multiple variables, or the value imposed by the randomization to depend upon the values of other variables. However, for the sake of simplifying the example, it will be assumed that the randomizations performed upon distinct variables are done independently, and that the value assigned in the randomization does not depend upon the values of other variables.

Following the basic framework of Strotz and Wold (1960), such a randomization can be represented in an LSEM by:

1. Replacing the equation original equation for each variable X that is randomized with a new equation $X :=\varepsilon'_X$, where ε'_X has the randomizing distribution; and
2. Setting the covariance between ε'_X and all other error terms to zero.

For example in Model 1, in order to manipulate the distribution of A to a normal distribution with mean 2 and variance 5, replace $A :=2 \times B+\varepsilon_A$ with $A :=\varepsilon\ '_A$, where ε'_A has mean 2 and variance 5. The new set of structural equations and assignments of values after the manipulation is: $A :=\varepsilon\ '_A$; $B :=\varepsilon_B$; $C :=.6B-0.4D+\varepsilon_C$; $D :=\varepsilon_D$; $E :=1{,}3C+\varepsilon_E$. The new parameters are $\boldsymbol{\Theta}'_1 = \langle\ \theta_{BC}=0.6,\ \theta_{DC}= -0.4,\ \theta_{CE}=1.3,\ \sigma'_A=\sqrt{5},\ \sigma_B=1,\ \sigma_C=1,\ \sigma_D=1,\ \sigma_E=1,\ \mu'_A=2,\ \mu_B=0,\ \mu_C=0,\ \mu_D=0,\ \mu_E=0\rangle$.

This new set of structural equations, and the probability distributions over the new error terms, determines a new joint probability distribution over the substantive variables. Note that because the new structural equation for A contains does not contain B, there is also a new path diagram (see Figure 2) that describes the manipulated population, in which there is no edge from B to A. The removal of this edge is needed for both the statistical and the causal interpretations of the graph of the manipulated model. In the manipulated structural equation model, A is independent of

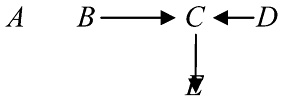

Figure 2 When A in G_1 is Manipulated

all of the other variables, and hence by the local directed Markov property no edge between A and any other (non-error) variable is needed. Also, in the manipulated structural equation model, A has no (non-error) causes, and hence under the causal interpretation, there is no edge between A and any other (non-error) variables.

Adapting the notation of Lauritzen (2001), if $f(A)$ is manipulated to a new probability distribution $f'(A)$, denote the new distribution by $f(A,B,C,D,E \| \{f'(A)\})$, where the '$\|$' notation denotes manipulation. In this example $f'(A) \sim \mathrm{N}(2,5)$. The special case in which every member of the population has been assigned the same value of a of A is denoted as $f(A,B,C,D,E \| A=a)$.

After manipulating A, marginal and conditional distributions can be formed in the usual ways. The marginal $f(B)$ after manipulating A to distribution $f'(A)$ is denoted by $f(B \| \{f'(A)\})$, and the manipulated distribution entailed by $<G_1,\Theta_1>$ is denoted as $f(B \| \{f'(A)\},<G_1,\Theta_1>)$. In this example, since $<G_1,\Theta_1>$ is hypothesized to be the true model, $f(B \| \{f'(A)\}) = f(B \| \{f'(A)\},<G_1,\Theta_1>) = f(B) \sim \mathrm{N}(0,1)$. Similarly, it is possible to first manipulate A to $f'(A)$, and then form the conditional distribution $f(B|A=a)$ after manipulating; this is denoted as $f(B|A=a \| \{f'(A)\})$.

The set of all joint distributions $f(A,B,C,D,E \| \mathbf{X},<G_1,\Theta_1>)$, where $\mathbf{X}$ ranges over every possible joint manipulation of the variables, is denoted by $\mathbf{M}(G_1, \Theta_1)$.

There are important differences between manipulating and conditioning. Both operations transform one probability distribution into a second probability distribution. In general, conditioning is relative to the set of values that are conditioned on, but manipulating is relative to a new probability distribution over the variables that are manipulated. Note, however, that it is possible to condition on a single value of a variable (e.g. $f(B|A=a)$), or to manipulate to a single value of a variable (e.g. $f(B \| A=a)$.

In addition, $f(A|B \in \mathbf{b})$, the distribution of A conditional on the value of B lying in the set of values $\mathbf{b}$ is a function of just the joint distribution of A and B (for conditioning sets that are not measure 0.) In contrast, $f(B \| \{f'(A)\})$ in a given population is a function not only of the joint distribution, but also of the true causal graph. For example, consider the graph G_2 of Figure 3. Let Model 2 be $<G_2,\Theta_2>$, where 2
$\boldsymbol{\Theta}_2 = \langle \theta'_{AB}=0.4,\ \theta_{BC}=0.6,\ \theta_{DC}=-0.4,\ \theta_{CE}=1.3,\ \sigma'_A=\sqrt{5}, \sigma'_B=\sqrt{0.2}, \sigma_C=1, \sigma_D=1, \sigma_E=1, \mu_A=0, \mu_B=0, \mu_C=0, \mu_D=0, \mu_E=0 \rangle$. $f(<G_1,\Theta_1>) = f(<G_2,\Theta_2>)$, i.e. the probability distributions are the same. However

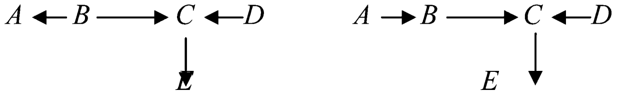

Figure 3 G_1 of Model 1; and G_2 of Model 2

$f(B\|f'(A),<G_1,\Theta_1>)\neq f(B\|f'(A),<G_2,\Theta_2>)$, because $f(B\|f'(A),$ $<G_2,\Theta_2>)\sim$N(0.8,1.0). Hence, $f(B\|\{f'A)\})$ depends upon which of G_1 or G_2 is the true causal graph.

The non-standard notation '$A :=2\times B+\varepsilon_A$' is used in order to emphasize that, according to Model 1, A is not just equal to $2\times B+\varepsilon_A$, but that it is being *assigned* the value $A :=2\times B+\varepsilon_A$. The difference between Model 1 and Model 2 shows up not in the sets of distributions that they represent, but in the effects of manipulations they entail. According to Model 1, manipulating A does not change the distribution of B, whereas according to Model 2 the distribution of B changes after manipulating A.

2.3 The relationship between the causal and statistical interpretations

The model $<G_1,\Theta_1>$ has now been used for two distinct purposes. First it determines a joint probability distribution over the substantive variables. Second, it determines the effects of manipulations. The effects of manipulations are not a function of the probability distribution alone. What assumption links these two distinct uses of the same model? The following assumption is a generalization of two assumptions: the immediate past screens off the present from the more distant past; and if X does not cause Y and Y does not cause X, then X and Y are independent conditional on their common causes.

> *Causal Markov Assumption:* Each variable is independent of its non-effects conditional on its direct causes.

In graphical terms, the Causal Markov Assumption states that in the population distribution, each variable is independent of its non-descendants and non-parents, conditional on its parents in the true causal graph. As long as the true causal graph is acyclic and the error terms do not cause each other and have no common causes, the Causal Markov Assumption is entailed for any structural equation model (linear or non-linear) by the weaker assumption that each error term is jointly independent of all of the other error terms.

This assumption presupposes that while the random variables of a unit in the population may causally interact, the units themselves are not causally interacting with each other. For example, if there is a population of people, at least to a high degree of approximation, Carol's exercise level affects the rate of her heartbeat, and Bob's exercise level affects the rate of his heartbeat, but Carol's exercise level does not affect the rate of Bob's heartbeat and Bob's exercise level does not affect the rate of Carol's heartbeat. On the other hand, Bob having measles may certainly affect Carol having measles and vice-versa. In that case, in order to apply the Causal Markov Assumption, the population has to be redefined to make a unit consist of a set of all of the people who may infect each other; a population of these units is then a population of sets of people.

It has often been pointed out (e.g. Yule 1926) that in seeming contradiction to the Causal Markov Assumption, two non-stationary time series can be correlated even if there are no causal connections between them. For example if bread prices in England and sea level in Venice are both increasing over time, they appear to be correlated even though there is no causal relation between them. Hoover (2003) argues that these apparent violations of the Causal Markov Assumption are due to the inappropriate use of the sample correlation coefficient as a test for association between trends in two different time series. When the appropriate statistical test for association between trends in two different time series is used, the apparent counterexamples are shown not to be violations of the Causal Markov Assumption.

The Causal Markov Assumption, together with the graphical representation of a manipulation as the breaking of all edges into the manipulated variables, entails the correctness of the rules for calculating the distribution after a manipulation.

3 PROBLEMS IN LSEM MODELING

The question investigated in this section is: are there reasonable assumptions under which it is possible to reliably estimate the effects of manipulations? In the course of answering this question, it will also be necessary to ask if there are reasonable assumptions under which it is possible to reliably find the true causal graph.

To be more specific, the following example will be used. Suppose that Model 1 in Figure 1 is the true causal model. It is not known what the true causal model is, but some sample data is available. For the purposes of illustration, assume as background knowledge that the correct model is an LSEM, and the correct model does not have latent variables, correlated errors, or cycles. (These unrealistic assumptions will be relaxed later, but they simplify the presentation of the basic ideas.) Further assume that some common measure of how well the data fits Model 1 [e.g. $p(\chi^2)$, or the Bayes Information Criterion] is high.[2] The goal is to answer the following three questions: What is the effect of manipulating A on B, i.e. $f(B \| \{f'(A)\})$? What is the effect of manipulating C on E, i.e. $f(E \| \{f'(C)\})$? What is the effect of manipulating B on C, i.e. $f(C \| \{f'B)\})$? In the context of LSEMs, these manipulated distributions are each described by two parameters, the mean and the variance of the affected variable.

3.1 One sense of 'reliable'

In the frequentist framework, a good (pointwise consistent) estimator of a quantity must approach the true value of the quantity in probability in the large sample limit regardless of what the true causal model is. Under reasonable assumptions, there are no good estimators in this strong sense either of causal graphs, or of the effect of manipulations, unless there is very strong background knowledge. (Note however that there are pointwise (and even uniform) consistent estimators of the effects of manipulations, *given* the correct causal graph.)

On the other hand, there are estimators of causal graphs or of the effects of manipulations that under some reasonable assumptions succeed except for models that are intuitively improbable. The frequentist framework does not provide any formal way of quantifying the improbability of models, but the Bayesian framework does. In the Bayesian framework, one method of point estimation of a quantity θ proceeds by:

1. Assigning a prior probability to each causal graph.
2. Assigning joint prior probabilities to the parameters Θ conditional on a given causal graph. (Assume this prior is assigned in such a way that conditional on the true causal graph, the posterior converges in the large sample limit to the correct parameter values with probability 1.)
3. Calculating the posterior probability of Θ (which is assumed to be a function of the posterior probabilities of the graphs and the graph parameter values.)
4. Turning the posterior probability over the mean and variance of the affected variable into a point estimate by returning the expected values of the parameters.

Unfortunately the strict Bayesian procedure is computationally infeasible for a number of reasons, including the fact that the number of graphs is superexponential in the number of variables. Note that such an estimator is a function not only of the data, but also of the prior probabilities. If the set of causal models (i.e. causal graph – probability distribution pairs) for which the estimator converges in probability to the correct value has a prior probability of 1, then say that it is *Bayes consistent* (with respect to the given set of priors.) Note that Bayes consistency is weaker than some other desirable properties in the Bayesian framework, such as minimizing mean squared error; however, it is not known how to find estimators with these stronger properties in a computationally feasible way. The procedures that I will describe for estimation are not themselves what an ideal Bayesian unrestricted by computational limitations would do, but under the assumptions described below they will satisfy a (slightly weakened version of) Bayes consistency.

4 WEAK BAYES CONSISTENCY

There are two major problems in constructing a Bayes consistent estimator of the effects of manipulations. They are both problems in which there is more than one causal theory compatible with a given probability distribution. I will argue that the first problem 'unfaithfulness' is unlikely to occur; and the second problem, 'distributional equivalence' can be solved by allowing estimators to return 'can't tell' in some (but not all) cases.

4.1 The problem of unfaithfulness

To illustrate the first kind of problem in estimating the effects of a manipulation, suppose for the moment that Model 1 is the (unknown) true model, and the only two possible models are Model 1 and Model 3 with the path diagrams G_1 and G_3 of Figure 4 respectively. (Subsequent sections will consider inference when any DAG model is possible.)

It can be shown that $\mathbf{P}(<G_1,\mathbf{\Theta_1}>)\subset\mathbf{P}(<G_3,\mathbf{\Theta_1}>)$. This entails that regardless of which $f(<G_1,\theta_1>)\in\mathbf{P}(<G_1,\mathbf{\Theta_1}>)$ is the true distribution, there is an alternative explanation equally compatible with the data, namely some $f(<G_3,\theta_3>)\in\mathbf{P}(<G_3,\mathbf{\Theta_3}>)$. Note also that $f(E\|\{f'(C)\},<G_1,\theta_1>)\neq f(E\|\{f'(C)\}, <G_3,\theta_3>)$, because according to Model 3, manipulating C has no effect on E, while according to Model 1, manipulating C has an effect on E. In general, regardless of what the true distribution $f(\mathbf{V})$ is, for *any* manipulation $f(X\|\{f'(\mathbf{Z})\})$, there are two models such that $f(\mathbf{V})=f(<G_A,\theta_A>)=f(<G_B,\theta_B>)$, but $f(X\|\{f'(\mathbf{Z})\},<G_A,\theta_A>)\neq f(X\|\{f'(\mathbf{Z})\}, <G_B,\theta_B>)$, and one of the manipulations had no effect on X. (This problem is closely related to the issue of choosing a matrix to premultiply a VAR model in order to produce a diagonal covariance matrix among the residuals.)

What are the priors under which there exist Bayes consistent estimators? If a DAG G does not entail that $\rho(X,Y|\mathbf{Z})=0$ for *all* legal values of the free parameters [i.e. $\rho(X,Y|\mathbf{Z}) \notin \mathbf{I}(G)$], nevertheless there may be *some* parameter values Θ such that $\rho(X,Y|\mathbf{Z})=0$ in $f(<G,\theta>)$. In that case say that $f(<G,\theta>)$ is *unfaithful* to G. For example, if $f(<G_3,\theta_3>)=f(<G_1,\theta_1>)$, then $\rho(B,D)=0$,

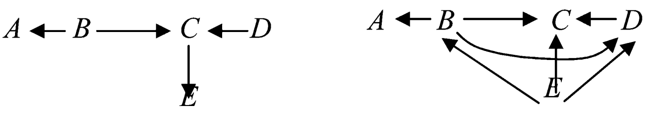

Figure 4 G_1 of Model 1, and G_4 of Model 4

because $\rho(B,D)=0$ is entailed by G_1. Because there is an edge between B and D in G_3, it follows that for any set $\mathbf{Z}$, $\rho(B,D|\mathbf{Z}) \notin \mathbf{I}(G_3)$. Hence if $f(<G_3,\Theta_3>)=f(<G_1,\Theta_1>)$, $\rho(B,D)=0$ but $\rho(B,D) \notin \mathbf{I}(G_3)$, and $f(<G_3,\Theta_3>)$ is unfaithful to G_3. (If $f(<G_3,\Theta_3>) \in \mathbf{P}(<G_1,\mathbf{\Theta_1}>)$ then there are also other zero partial correlations that are not in $\mathbf{I}(G_3)$ as well.

Consider the subgraph of G_3 containing just the vertices B, D, and E. Suppose that the variances of B, D, and E are fixed at 1, and that B represents *Total Income*, E represents *Tax Rates*, and D represents *Tax Revenues*. In that case the total correlation between *Total Income* and *Tax Revenues* is the sum of two terms. The first term is due to the direct positive effect of *Total Income* on *Tax Revenues* (θ_{BD}) and is positive. The second term is due to the product of the direct negative effect of *Tax Rates* on *Total Income* (θ_{EB}) and the direct positive effect of *Tax Rates* on *Tax Revenue* (θ_{ED}) and is negative. If the two terms exactly cancel each other (i.e. $\theta_{BD}=-\theta_{EB,} \times \theta_{ED}$) there is a zero correlation between *Total Income* and *Tax Revenues*.

This algebraic constraint (the cancellation of the two terms) defines a surface of unfaithfulness in the parameter space. Every set of parameter values for which $f(<G_3,\Theta_3>) \in \mathbf{P}(<G_1,\mathbf{\Theta_1}>)$ lies on the surface of unfaithfulness. This polynomial in the free parameters is a two-dimensional surface in the three dimensions of the space of parameters (assuming the variances are fixed). Figure 5 shows the plane in which $\theta_{BD}=0$, and the two-dimensional surface on which $\theta_{BD}=-\theta_{EB,} \times \theta_{ED}$. (The set of legal parameter values actually extends beyond the box shown in Figure 5 because θ_{BD} and θ_{ED} range from $-\infty$ to ∞, and there are some additional constraints on the parameters that place joint limitations on the parameters that are not shown. In order for Model 3 to unfaithfully represent a distribution in $\mathbf{P}(<G_1,\mathbf{\Theta_1}>)$, other algebraic constraints, which are not shown, must be satisfied as well).

Note that *any* value of θ_{BD} is compatible with $\rho(B,D)=0$ because each value of θ_{BD} occurs somewhere on the surface of unfaithful parameters. So, if E were unobserved, any possible direct effect of B on D would be compatible with the observed correlation $\rho(B,D)=0$.

In general, for parametric models, the set of parameters associated with unfaithful distributions is of lower dimension than the full parameter space, and hence is of Lebesgue measure 0. If the only two choices were between Model 1 and Model 3, and the population distribution was faithful to G_1 (and hence unfaithful to G_3) it would follow that for any prior distribution that assigns non-zero probability to each graph, and a zero probability to lower dimension subspaces of the parameter space, with probability 1 in the large sample limit the posterior probability of Model 1 approaches 1 and the posterior probability of Model 3 approaches 0. Assuming that each graph is assigned a non-zero probability, and (temporarily) assuming the only two

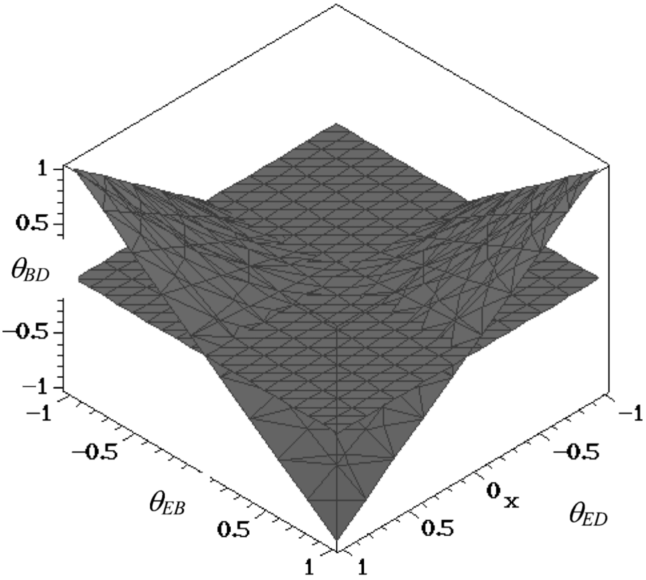

Figure 5 Surface of Unfaithfulness

DAGs with positive probability are G_1 and G_3, the following assumption is sufficient to guarantee the existence of Bayes consistent estimators:

> *Bayesian Causal Faithfulness Assumption:* If G is the true causal graph, and G does not entail **X** is independent of **Y** conditional on **Z** (i.e. **X** independent of **Y** conditional on $\mathbf{Z} \notin \mathbf{I}(G)$) the set of parameter values $\boldsymbol{\Theta}$ such that **X** is independent of **Y** conditional on **Z** in $f(G,\Theta)$ has prior probability zero.

In the example described above, this entails assigning the set of parameter values in Model 3 such that $\theta_{BD}=-\theta_{EB,} \times \theta_{ED}$ (among others) a zero prior probability.[3] Given the Causal Markov Assumption and the Bayesian Causal Faithfulness Assumption, with probability 1, **X** and **Y** are independent conditional on **Z** in $f(\mathbf{V})$ if and only if **X** is d-separated from **Y** conditional on **Z** in the true causal graph.

4.1.1 Faithfulness and common methods of LSEM selection

The Bayesian Causal Faithfulness Assumption is implicit in several different common methods of LSEM selection. One method of LSEM selection is to use background knowledge to construct the model, and a χ^2 statistical test to determine whether to accept the model. Nevertheless, it is possible that Model 3 is true but that because $f(<G_3,\Theta_3>)=f(<G_1,\Theta_1>)$, the statistical test would mistakenly select Model 1 as the true LSEM. (Note that the test is correct about the sample fitting Model 1; it is only incorrect if the test is used to select Model 1 as a *causal* model.) Use of such a test to select LSEMs implicitly assumes that such violations of faithfulness are of low probability (which we are approximating by assuming zero probability.)

A second method of LSEM selection is to assign a score to each model, and choose the model with the higher score. Under a model score such as the Bayesian Information Criterion (BIC) in the large sample limit the probability of selecting Model 1 over Model 3 is equal to 1. This is because the Bayesian Information Criterion is a penalized maximum likelihood score that rewards a model for assigning a high likelihood to the data (under the maximum likelihood estimate of the values of the free parameters), and penalizes a model for being complex (which for causal DAG models without latent variables can be measured in terms of the number of free parameters in the model.) For such models, the Bayes Information Criterion is also a good approximation to the posterior probability in the large sample limit.

Model 3 is more complex (has a higher dimension) than Model 1. In the large sample limit, the penalty for the higher dimension of Model 3 will increase fast enough to ensure with probability 1 that Model 1 will be preferred, given that $f(<G_1,\Theta_1>) \in \mathbf{P}(<G_1,\mathbf{\Theta_1}>)$ is the true distribution. Nevertheless, it is possible that Model 3 is true but that because $f(<G_3,\Theta_3>)=f(<G_1,\Theta_1>)$, the BIC would mistakenly select Model 1. Use of a score such as BIC to select causal models implicitly assumes that such violations of faithfulness are of low probability (which is approximated by assuming zero probability.)

A common practice for variable selection in constructing causal models for a variable T is to regress T on all other measured variables, and to remove the variables that have insignificant regression coefficients. The implicit justification for this procedure is that if the regression coefficient for X is insignificant, then it is probable that the effect of X on T is small. Assuming linearity, the regression coefficient for X when T is regressed on all of the other observed variables is equal to zero if and only if the partial correlation between T and X conditional on all the other observed variables is also zero. So this practice also implicitly assumes, at least approximately, an instance of the Bayesian Causal Faithfulness Assumption.

4.1.2 Faithfulness and priors

In general, for parametric models, the set of parameters associated with unfaithful distributions is of lower dimension than the full parameter space, and hence is of Lebesgue measure 0. Any Bayesian who assigns a prior absolutely continuous with Lebesgue measure (as all of the typical priors are) to the linear coefficients in an LSEM is adopting a prior that satisfies the Bayesian Causal Faithfulness Assumption. Of course, a Bayesian is free to assign non-standard priors that violate the Bayesian Causal Faithfulness

Assumption, and there are circumstances where a prior obeying the Bayesian Causal Faithfulness Assumption is not appropriate. So the status of the Bayesian Faithfulness Assumption is that it functions as a default that can be overridden by specific knowledge about a system.

One kind of case in which the Bayesian Faithfulness Assumption should not be made is if there are deterministic relationships between measured variables. If the true causal graph is $X \rightarrow Y \rightarrow Z$, and $X = Y = Z$, then X and Y are independent conditional on Z (because X is a function of Y), even though X and Y are not d-separated conditional on Z in the true causal graph. Theoretically it is possible to detect when there are deterministic relations among the observed variables, in which case the Bayesian Faithfulness Assumption should not be made. In practice it may be difficult to detect that some variable is a function of a large set of other variables.

A second kind of case in which the Bayesian Faithfulness Assumption should not be made are cases in which parameters may be deliberately chosen in such a way as to violate the Bayesian Faithfulness Assumption. Hoover (2001) describes an example where policy makers adopt a rule for minimizing the variance of GDP at time t by controlling the money stock. The parameters relating money stock at time t to money stock and GDP at t-1 that minimize the variance of GDP at t also lead to violations of faithfulness, in which both GDP at t-1 and money stock at t-1 are uncorrelated with GDP at t.

Even given the Bayesian Faithfulness Assumption, there are no pointwise consistent estimators of either the correct causal graph or of the effects of a manipulation, because any estimator is sometimes wrong (albeit on a set of measure 0), e.g. when $f(<G_3,\Theta_3>)=f(<G_1,\Theta_1>)$ and there is no way to reliably decide which model is correct.

The Bayesian Faithfulness Assumption excludes a non-zero probability for parameters Θ_3 such that $f(<G_3,\Theta_3>)=f(<G_1,\Theta_1>)$, but it does not exclude a high probability for 'near-unfaithfulness', i.e. that $f(<G_3,\Theta_3>)$ is arbitrarily close to $f(<G_1,\Theta_1>)$. In that case, Model 3 still predicts an effect of manipulating C on E that is zero, and hence far away from the prediction of Model 1. For the usual model selection methods, detecting that Model 3 is correct if it is near-unfaithful requires very large sample sizes. An open research question is whether there are plausible stronger versions of the Bayesian Faithfulness Assumption that would make near-unfaithfulness unlikely. (For a discussion of near-unfaithfulness in the frequentist frameworks, see Spirtes, Glymour and Scheines 2000; and Robins *et al.* 2003.)

4.2 Distributional equivalence

It can be shown that Model 1 and Model 2 in Figure 6 represent exactly the same set of probability distributions, i.e. $\mathbf{P}(<G_1,\mathbf{\Theta_1}>)=\mathbf{P}(<G_2,\mathbf{\Theta_2}>)$. In that case say that $<G_1,\boldsymbol{\Theta}_{\mathbf{1}}>$ and $<G_2,\boldsymbol{\Theta}_{\mathbf{2}}>$ are *distributionally equivalent*. Whether two models are distributionally equivalent depends not only on the graphs in the models, but also on the parameterization families of the models (e.g. multivariate normal). A set of models that are all distributionally equivalent to each other is a *distributional equivalence class*. If the graphs are all restricted to be DAGs, then they form a *DAG distributional equivalence class*.

G_1 and G_2 are *conditional independence equivalent* if and only if $\mathbf{I}(G_1)=\mathbf{I}(G_2)$. In contrast to distributional equivalence, conditional independence equivalence depends only upon the graphs of the models, and not on the parameterization families. Conditional independence equivalence of G_1 and G_2 is a necessary, but not always sufficient condition for the distributional equivalence of $<G_1,\boldsymbol{\Theta}_{\mathbf{A}}>$ and $<G_2,\boldsymbol{\Theta}_{\mathbf{B}}>$. In the case of multivariate normal or discrete distributions without latent variables, conditional independence equivalence does entail distributional equivalence. A set of graphs that are all conditional independence equivalent to each other is a *conditional independence equivalence class*. If the graphs are all restricted to be DAGs, then they form a *DAG conditional independence equivalence class*. G_1 and G_2 are conditional independence equivalent, and form a DAG conditional independence equivalence class.

$<G_1,\boldsymbol{\Theta}_{\mathbf{1}}>$ and $<G_2,\boldsymbol{\Theta}_{\mathbf{2}}>$ of Figure 6 are different models only in the sense that they differ in their predictions of the effects of some manipulations, i.e. $\mathbf{M}(<G_1,\boldsymbol{\Theta}_{\mathbf{1}}>)\neq\mathbf{M}(<G_2,\boldsymbol{\Theta}_{\mathbf{2}}>)$. For example, Model 1 predicts that manipulating A has no effect on B; Model 2 predicts that manipulating A has an effect on B. Because they represent the same set of probability distributions, without further background knowledge no reliable statistical inference from the data can distinguish between them. On the usual scores of models [e.g. $p(\chi^2)$, BIC, etc.] they receive the same score on every data set.

Suppose that it were known that either Model 1 or Model 2 is correct, but it is not known which of Model 1 or Model 2 is correct. In that case (as shown in more detail below) $P(C\|P'(B)),<G_1,\boldsymbol{\Theta}_{\mathbf{1}}>)=P(C\|P'(B),<G_2,\boldsymbol{\Theta}_{\mathbf{2}}>)$, and $P(E\|P'(C)),<G_1,\boldsymbol{\Theta}_{\mathbf{1}}>)=P(E\|P'(C),<G_2,\boldsymbol{\Theta}_{\mathbf{2}}>)$. So if it were known that either Model 1 or Model 2 is correct, there are pointwise consistent estimators of $P(C\|P'(B))$ and $P(E\|P'(C))$, even without knowing *which* of Model 1 or Model 2 is correct.

On the other hand, in section 2.2 it was shown that $f(B\|f'(A),<G_1,\Theta_1>)\sim N(0,1)$, but $f(B\|f'(A),<G_2,\Theta_2>)\sim N(0.8,1.0)$. This is because G_1

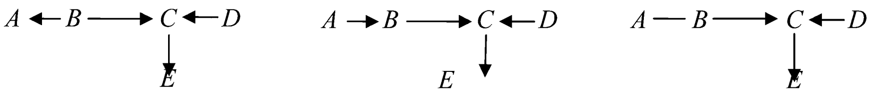

Figure 6 G_1 of Model 1; G_2 of Model 2; and P_1 (pattern for G_1 and G_2)

predicts that manipulating A has no effect on B, while G_2 predicts that manipulating A has an effect on B. Since without further information there is no way to determine which of these two estimates is correct, there is no Bayes (or pointwise) consistent estimator of $f(B \| f'(A))$. For Bayes estimators, in the limit the posterior probability will peak around two different estimates of the mean. In this example, it would be possible to output both estimates, but in cases with large numbers of DAGs in the DAG distributional equivalence class, or when the possibility of latent common causes is allowed, this is not computationally feasible.

In cases where it is not computationally feasible to output all of the different estimates, another correct response is to output 'can't tell'. Say that an estimator is *weak Bayes consistent* if on a set of causal graph-parameter pairs of prior probability 1 in the large sample limit the estimator approaches the true value in probability, or it outputs 'can't tell'. (There is an analogous weakening of pointwise consistent estimators to weak pointwise consistent estimators.) Of course, if an estimator outputs 'can't tell' in every case, it would be a weak Bayes consistent estimator, but uninteresting. Say that an estimator is *non-trivial* if there are causal graph–parameter pairs such that with probability 1 in the large sample limit, the estimator outputs a numerical estimate for those graph-parameter pairs. There are non-trivial weak Bayes consistent estimators of the effects of manipulations, given the Causal Markov and Causal Faithfulness Assumptions.

In general, a necessary condition for the existence of a weak Bayes consistent estimator of the effect of a manipulation is that every model that is distributionally equivalent to the true model agrees on the predicted effect of the manipulation. Hence, in order to determine whether the necessary condition is satisfied for the effect of manipulating B on C, it is necessary to know whether the complete set of LSEM DAG models that are distributionally equivalent to Model 1 agree about their predictions of the effects of manipulating B on C.

Unfortunately, there is no known computationally feasible algorithm for determining when two models are distributionally equivalent for arbitrary parameterizations of a DAG. In part, this is because whether two models $<G_1,\mathbf{\Theta_A}>$ and $<G_2,\mathbf{\Theta_B}>$ are distributionally equivalent depends not only upon G_1 and G_2, but also on the parameterizations $\mathbf{\Theta_A}$ and $\mathbf{\Theta_B}$. However, if G_1 and G_2 are DAGs, there is a known general computationally feasible algorithm for testing when two models $<G_1,\mathbf{\Theta_A}>$ and $<G_2,\mathbf{\Theta_B}>$ are conditional independence equivalent; conditional independence equivalence depends only upon G_1 and G_2, and not on the parameterizations. This test forms the basis of a general computationally feasible (for graphs in which no variable has a large number of causes) non-trivial weak Bayes consistent estimator of the effects of manipulations. The price that one pays for constructing the estimator using a test of conditional independence equivalence, rather than distributional equivalence, is that for some parameterizations, the former is less informative (i.e. outputs 'can't tell' more often) than the latter. However, as long as the variables are either multivariate normal or all discrete, the former estimator is not less informative than the latter estimator.

The reason that a graphical test of conditional independence equivalence is sufficient for the purpose of constructing a non-trivial weak Bayes consistent estimator of the effects of manipulations, is that the set of DAGs that are all conditional independence equivalent is a superset of the set of DAGs that are in models that are distributionally equivalent for a given parameterization. Hence, if all of the DAGs that are conditional independent equivalent to the DAG of a given model agree on their predictions about the effects of a particular manipulation, then so do all of the models that are distributionally equivalent to a given model. Theorem 1 (that can be described as the Observational Equivalence Theorem) was proved in Verma and Pearl (1990).

> *Theorem 1 (Observational Equivalence Theorem):* Two directed acyclic graphs are conditional independence equivalent if and only if they contain the same vertices, the same adjacencies, and the same unshielded colliders.

Theorem 1 entails that the set consisting of G_1 and G_2 is a DAG conditional independence equivalence class.

4.3 Patterns: features common to a DAG conditional independence equivalence class

Theorem 1 is also the basis of a simple representation (called a 'pattern' in Verma and Pearl, 1990) of a DAG conditional independence equivalence class. Patterns can be used to determine which predicted effects of a manipulation are the same in every member of a DAG conditional independence equivalence class and which are not.

A 'pattern' P represents a DAG conditional independence equivalence class **X** if and only if:

1. P contains the same adjacencies as each of the DAGs in **X**;
2. each edge in P is oriented as $X \to Z$ if and only if the edge is oriented as $X \to Z$ in every DAG in **X**, and as X–Z otherwise.

Meek (1995), Andersson *et al.* (1995), and Chickering (1995) show how to 3 generate a pattern from a DAG. The pattern P_1 for the DAG conditional independence equivalence class containing G_1 of Model 1 is shown in Figure 6. It contains the same adjacencies as G_1, and the edges are the same except that the edge between A and B is undirected in the pattern, because it is oriented as $A \leftarrow B$ in G_1, and oriented as $A \rightarrow B$ in G_2.

4.4 Estimating the effects of a manipulation from a given pattern

Suppose that pattern P_1 is given. There is a general procedure for using the pattern to determine whether there are pointwise or Bayes consistent estimators of the effect of a manipulation, and if so, what the estimator is. Here I will simply illustrate the procedure for two simple cases.

Consider first the problem of estimating $f(E\|\{f'(C)\})$ given P_1, where $f'(C)$~N(2,1). Because manipulated distributions are just probability distributions, it follows from the chain rule that:

$$f(E\|f'(C)) = \int_{-\infty}^{\infty} f(E|C\|\{f'(C)\})f'(C\|\{f'(C)\})dC \tag{1}$$

where $f'(C\|f'(C))$ is the distribution of C after manipulating the distribution of C to f'(C); this is $f'(C)$ by definition.

Because all of the paths from C to E in P_1 that do not contain colliders are out of C, it is possible to prove that $f(E|C\|\{f'(C)\})=f(E|C)$ for every member of the equivalence class represented by P_1, i.e. that the distribution of E conditional on C is the same before and after the manipulation. For both G_1 and G_2, $f(E|C=c\|\{f'(C)\})=f(E|C=c)$~N(1.3$c$,1) when $f'(C)$~N(2,1). It follows from (1) then that:

$$f(E\|f'(C)) = \int_{-\infty}^{\infty} f(E|C\|f'(C))f(C\|f'(C))dC = \int_{-\infty}^{\infty} f(E|C)f'(C)dC \tag{2}$$

$f(E|C)$ can be estimated from the observed distribution in the usual way. $f'(C)$ is given. Substituting the estimate for $f(E|C)$ and the given $f'(C)$ into equation 1 provides a pointwise consistent estimator of $f(E\|\{f'(C)\})$, which in this example is N(2.6,2.69).

$f(C\|\{f'(B)\})$ can be estimated in an analogous way.

Consider next the problem of estimating $f(B\|\{f'(A)\})$. It is clear from the undirected edge between A and B in the pattern that some members of the conditional independence equivalence class (in this case G_1) predict that manipulating A has no effect on B, while other members of the conditional independence equivalence class (in this case G_2) predict that manipulating A has an effect on B. Since without further information there is no way to determine which of these two estimates is correct, there is no pointwise or Bayes consistent estimator of $f(B\|\{f'(A)\})$ even given P_1. In this case a correct response of an estimator is 'can't tell'. Once the possibility of 'can't tell' as the output of an estimator is allowed, there are nontrivial weak pointwise or Bayes consistent estimators of the effects of manipulations, given P_1.

4.5 Searching the space of DAG conditional independence equivalence classes

Given a pattern, there are nontrivial weak pointwise or Bayes consistent estimators of the effects of manipulations. If there is a Bayes consistent estimator of patterns, then the pattern estimator and the estimator of the effect of a manipulation given a pattern can be put together to provide a nontrivial weak Bayes consistent estimator of the effects of manipulations. There are two major different approaches to estimating patterns: constraint based search, described below, and score based search (Chickering 2002). The Bayesian Causal Faithfulness Assumption is taken as given for both kinds of searches.

4.5.1 Constraint-based search

A constraint-based search attempts to find the pattern that most closely entails all and only the conditional independencies judged to hold in the population. It has an adjacency phase in which the adjacencies are determined, and an orientation phase in which as many edges as possible are directed.

The adjacency phase is based on the following two theorems, where **Parents**(G,A) is the set of parents of A in G.

> *Theorem 2:* If A and B are d-separated conditional on any subset $\mathbf{Z}$ in DAG G, then A and B are not adjacent in G.
>
> *Theorem 3:* A and B are not adjacent in DAG G if and only if A and B are d-separated conditional on **Parents**(G,A) or **Parents**(G,B) in G.

The algorithm is stated below. First the algorithm is illustrated in Figure 7, supposing that P_1 is the true unknown pattern, but that it is possible to perform tests of conditional independence on the observed variables. First, the algorithm starts with a graph in which every pair of vertices is connected by an undirected edge, as in (i). Then for each pair of vertices X and Y, the algorithm tests whether they are independent (which under the Causal Markov and Bayesian Causal Faithfulness Assumptions amounts to testing whether they are d-separated conditional on the empty set in the true causal graph). If they are the algorithm removes the edge, and otherwise the algorithm leave the edge in. This is illustrated in (ii). This step is justified by Theorem 2. In (iii), for each pair of vertices that are still adjacent, such as A and C, the algorithm tests whether they are independent conditional on any vertex adjacent to A or adjacent to C. For example, A and C are independent conditional on $\{B\}$, which is adjacent to A, so the algorithm

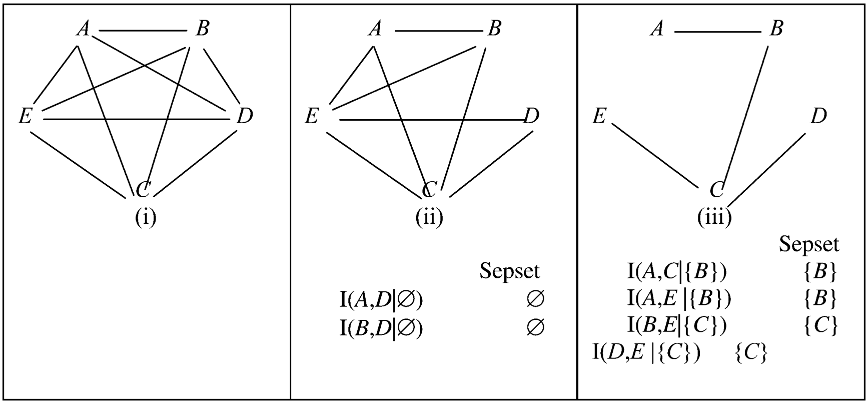

Figure 7 Constraint based search, where correct pattern is P_1

removes the edge between *A* and *C*. Again, this is justified by Theorem 2. After stage (iii), all of the adjacencies that are left are correct, but the algorithm does not know this yet. In the next stage of the algorithm (not shown in Figure 7) for each pair of vertices such as *C* and *E* that are still adjacent, the algorithm tests whether they are independent conditional on any pair of vertices adjacent to *C* (and not containing *E*) or pair of vertices adjacent to *E* (and not containing *C*). There is only one such pair of vertices, namely *B* and *D*. So the algorithm tests whether *C* and *E* are independent conditional on $\{B,D\}$, and it finds that they are not. Similarly, the algorithm tests whether *C* and *D* are independent conditional on $\{B,E\}$, and whether *C* and *B* are independent conditional on $\{D,E\}$. Note that in the case of *A* and *B* there is no pair of vertices (not containing *A* or *B*) that are both adjacent to *A* or both adjacent to *B*, so this stage of the algorithm never tests whether *A* and *B* are independent conditional on any subset of the other variables. Finally, since there is no triple of variables all adjacent to one of the endpoints of any remaining edge (and not containing the other endpoint), the adjacency phase of the algorithm halts.

By Theorem 2 and the Bayesian Causal Faithfulness Assumption, every edge that has been removed is not in the true pattern. Hence, at each stage of the algorithm **Parents**(*G*,*X*) is a subset of the vertices still adjacent to *X*. If *X* and *Y* are still adjacent when the adjacency phase of the search ends, whether *X* and *Y* are independent conditional on **Parents**$(G,X)\backslash\{Y\}^4$ or conditional on **Parents**$(G,Y)\backslash\{X\}$ has been tested, because the algorithm tests whether *X* and *Y* are independent conditional on *every* subset of variables adjacent to *X* (excluding *Y*), which includes **Parents**$(G,X)\backslash\{Y\}$ as a subset; similarly the algorithm has tested whether *X* and *Y* are independent conditional on *every* subset of variables adjacent to *Y* (excluding *X*) which includes **Parents**$(G,Y)\backslash\{X\}$. Hence by Theorem 3, the algorithm has removed every edge that is not in the true pattern.

The adjacency phase of the algorithm is stated more formally below. Let **Adjacencies**(C,A) be the set of vertices adjacent to A in graph C. (In the algorithm, the graph C is continually updated, so **Adjacencies**(C,A) is constantly changing as the algorithm progresses.)

Adjacency Phase of PC Algorithm

```
Form an undirected graph $C$ in which every pair of vertices in $\mathbf{V}$ is adjacent.
$n := 0$.
repeat
    repeat
        Select an ordered pair of variables $X$ and $Y$ that are adjacent in $C$ such that Adjacencies($C$,$X$)$\setminus\{Y\}$ has cardinality greater than or equal to $n$, and a subset $\mathbf{S}$ of Adjacencies($C$,$X$)$\setminus\{Y\}$ of cardinality $n$, and if $X$ and $Y$ are independent conditional on $\mathbf{S}$ delete edge $X - Y$ from $C$ and record $\mathbf{S}$ in Sepset($X$,$Y$) and Sepset($Y$,$X$);
    until all ordered pairs of adjacent variables $X$ and $Y$ such that Adjacencies($C$,$X$)$\setminus\{Y\}$ has cardinality greater than or equal to $n$ and all subsets $\mathbf{S}$ of Adjacencies($C$,$X$)$\setminus\{Y\}$ of cardinality $n$ have been tested for conditional independence;
    $n := n+1$;
until for each ordered pair of adjacent vertices $X$, $Y$, Adjacencies($C$,$X$)$\setminus\{Y\}$ is of cardinality less than $n$.
```

After the adjacency phase of the algorithm, the orientation phase of the algorithm is performed. The correctness of the orientation phase of the algorithm is based on the following theorem, and is illustrated in Figure 8.

Theorem 4: If in a DAG G, A and B are adjacent, B and C are adjacent, but A and C are not adjacent, either B is in every subset of variables $\mathbf{Z}$ such that A and C are d-separated conditional on $\mathbf{Z}$, in which case $<A,B,C>$ is not a collider, or B is in no subset of variables $\mathbf{Z}$ such A and C are d-separated conditional on $\mathbf{Z}$, in which case $<A,B,C>$ is a collider.

The first phase of the orientation algorithm is illustrated in Figure 8(i). The boldfaced lines indicate which orientations are added in that phase of the algorithm. The first phase of the algorithm looks for triples of variables such that X and Y are adjacent, Y and Z are adjacent, and X and Z are not adjacent. For example, the edge between B and D was removed by the algorithm because B and D are independent conditional on **Sepset**(B,D)$=\phi$. $C \notin$ **Sepset**(B,D), so by Theorem 4 (and the Causal Markov and Bayesian Causal Faithfulness Assumptions), C is not a member of any set of vertices $\mathbf{R}$ such that B and D are d-separated conditional on $\mathbf{R}$. Then by Theorem 4 the edges are oriented as $B \rightarrow C \leftarrow D$. On the other hand, because $B \in$ **Sepset**(A,C), then by Theorem 4, B is a member of every set $\mathbf{R}$ such

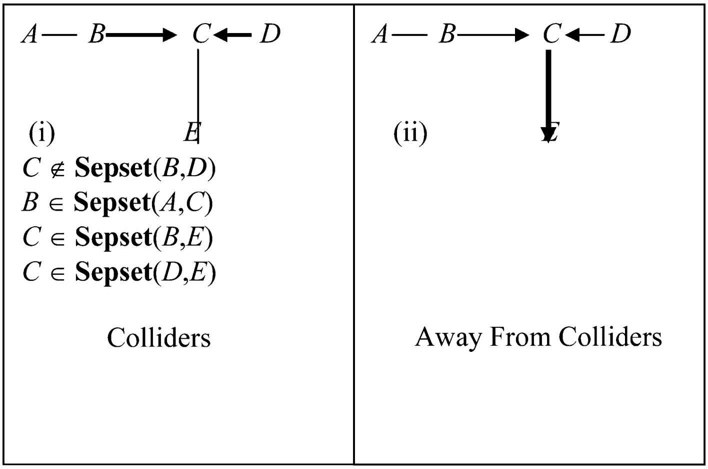

Figure 8 Orientation phase of PC algorithm, assuming true pattern is P_1

that A and C are d-separated conditional on **R**. By Theorem 4, the triple $<A,B,C>$ is not a collider in G, so the orientation $A–B\rightarrow C$ in the pattern is left unchanged.

The next phase of the orientation algorithms is shown in Figure 8(ii). The orientations so far include $B\rightarrow C–E$. The edges between B and C, and C and E do not collide at C, because otherwise they would have been oriented as a collider in the previous phase. The only way that they do not collide at C is if the $C–E$ edge is oriented as $C\rightarrow E$.

The orientation phase of the PC algorithm is stated more formally below. The last two orientation rules (Away from Cycles, and Double Triangle) are not used in the example, but are sound because if the edges were oriented in ways that violated the rules, there would be a directed cycle in the pattern, which would imply a directed cycle in the graph (which in this section is assumed to be impossible). Meek (1995) proved that the orientation rules are complete (i.e. every edge that has the same orientation in every member of a DAG conditional independence equivalence class is oriented by these rules.)

Orientation Phase of PC Algorithm

```
For each triple of vertices $X$, $Y$, $Z$ such that the pair $X$, $Y$ and the pair $Y$, $Z$ are each adjacent in graph $C$ but the pair $X$, $Z$ are not adjacent in $C$, orient $X–Y–Z$ as $X\rightarrow Y\leftarrow Z$ if and only if $Y$ is not in Sepset($X$,$Z$).
repeat
    Away from colliders: If $A\rightarrow B–C$, and $A$ and $C$ are not adjacent, then orient as $B\rightarrow C$.
    Away from cycles: If $A\rightarrow B\rightarrow C$ and $A–C$, then orient as $A\rightarrow C$.
```

Double Triangle: If $A \rightarrow B \leftarrow C$, A and C are not adjacent, $A - D - C$, and there is an edge $B - D$, orient $B - D$ as $D \rightarrow B$.

until no more edges can be oriented.

The tests of conditional independence can be performed in the usual way. Such tests require specifying a significance level for the test, which is a user-specified parameter of the algorithm. Because the PC algorithm performs a sequence of tests without adjustment, the significance level does not represent any (easily calculable) statistical feature of the output, but should only be understood as a parameter used to guide the search. Nevertheless, under the Causal Markov Assumption, and the Bayesian Causal Faithfulness Assumption, the PC algorithm is a Bayes consistent estimator of the true pattern. The PC algorithm together with the algorithm for estimating the effects of manipulations from patterns forms a non-trivial weak Bayes consistent estimator of the effects of manipulations.

5 LATENT COMMON CAUSES

The same general approach towards inference of the effects of manipulations can be taken even when there is the possibility of latent common causes. When the possibility of latent common causes is admitted the following modifications to the four basic parts of constructing a non-trivial weak Bayes consistent estimator of the effects of a manipulation must be made.

1. The Causal Markov Assumption and the Bayesian Causal Faithfulness Assumption do not need to be modified.
2. For the purposes of inference, what is of interest is marginal conditional independence equivalence, where the marginal distribution is over the observed variables. (Conditional independence relations that involve unobserved variables cannot be tested and are of no use in inference based upon observed conditional independence relations.) A kind of graph, called a 'partial ancestral graph', represents (some, but not all) features common to every model in a DAG marginal conditional independence equivalence class. It requires several more kinds of edges than a pattern does, including double-headed arrows $X \leftrightarrow Y$, which indicate the presence of a latent common cause between X and Y.
3. There is a known extension of the constraint based PC algorithm (the Fast Causal Inference Algorithm, Spirtes *et al.* 1993) that 4 searches over the space of marginal conditional independence classes, and outputs a partial ancestral graph. It is a Bayes consistent estimator of the correct marginal DAG conditional independence equivalence class, and of some of the features the DAGs in the class have in common. The Fast Causal Inference Algorithm is more

complicated and slower than the PC algorithm, but is computationally feasible for around 100 variables, if the true partial ancestral graph is sparse.

4. There is a known computationally feasible (if the partial ancestral graph is sparse) weak pointwise consistent algorithm for estimating the effects of a manipulation from a partial ancestral graph (or outputting 'can't tell'). It is slower and more complicated than the corresponding algorithm for estimating the effects of manipulations from a pattern, and is a less informative estimator. It is not known whether it is complete. See Spirtes *et al.* (1993, 1999).

Consider again the problems of estimating $f(B\|\{f'(A)\})$, $f(E\|\{f'(C)\})$ and $f(C\|\{f'(B)\})$, only now allow the possibility that there are latent common causes. Suppose that the DAG marginal conditional independence equivalence class is given. Now the DAG marginal conditional independence equivalence class containing G_1 also contains G_4, as well as an infinite number of other DAGs that are not shown. As in the case with no latent common causes, any non-trivial weak pointwise consistent estimator of $f(B\|\{f'(A)\})$ outputs 'can't tell' because $<G_1,\theta_1>$ and $<G_2,\theta_2>$ make different predictions about $f(B\|\{f'(A)\})$. As in the case with no latent common causes, there is a non-trivial weak pointwise consistent estimator of $f(E\|\{f'(C)\})$ that outputs a numerical estimate because it can be shown that all of the members of the DAG marginal conditional independence equivalence class containing $<G_1,\Theta_1>$ agree on their predictions about $f(E\|\{f'(C)\})$. In contrast to the case with no latent common causes, however, any non-trivial weak pointwise consistent estimator of $f(C\|\{f'(B)\})$ outputs 'can't tell' because while $<G_1,\theta_1>$ and $<G_2,\theta_2>$ make the same predictions about $f(C\|\{f'(B)\})$, $<G_4,\theta_4>$ makes a different prediction (that manipulating B has no effect on C) about $f(C\|\{f'(B)\})$ than either $<G_1,\theta_1>$ or $<G_2,\theta_2>$.

6 EXTENSIONS TO CYCLES AND VAR MODELS

Systems in some kinds of equilibrium can be represented by cyclic directed graphs. The situation with respect to directed graphs with cycles is similar in spirit to the cases of DAGs with or without latent variables. Details about how to represent a conditional independence equivalence class of graphs that may contain cycles, and how to search the conditional independence equivalence class of graphs that may contain cycles are given in Richardson (1996). One important difference between the theory of graphs with cycles and those without cycles is that the former applies only to multivariate normal or discrete distributions that have reached equilibrium. Non-linear relationships between continuous variables require a much more radical revision to the theory, introduce many more 'can't tell' answers, and the best way to handle these cases is an open question.

Figure 9 G_1 of Model 1; and G_4 of Model 4 11

Swanson and Granger (1997), Bessler and Lee (2002), Demiralp and Hoover (2003), and Moneta (2003) used various modifications of the PC algorithm to search for and selecting causal orderings among contemporaneous variables in VAR models.

7 FINITE SAMPLE SIZES

How well does the PC algorithm perform at finite sample sizes? With the appropriate Causal Markov and Bayesian Causal Faithfulness assumptions, the non-trivial weak Bayes consistency of the estimators described in this article applies only to a limited class of distributional families for which tests of conditional independence are available (including multivariate Normal or multinomial) and to a limited class of models (DAGs with or without latent variables, linear or discrete cyclic graphs and VAR models). In addition, they do not always use all available background knowledge (e.g. parameter equality constraints.) How well an estimator performs on actual data depends upon at least five factors:

1. The correctness of the background knowledge input to the algorithm;
2. Whether the Causal Markov Assumption holds;
3. Whether the true model is near-unfaithful;
4. Whether the distributional assumptions made by the statistical tests of conditional independence hold;
5. The power of the conditional independence tests used by the estimators (which depends in part on the sample size, and the number of variables in the conditioning set).

Each of these factors may negatively affect the output in particular cases. Hence the output of the estimators described in this chapter should be subjected to further tests wherever possible. However, the problem is made even more difficult because even under the Bayesian Causal Faithfulness Assumption, for computational reasons, it is not known how to probabilistically bound the size of errors. It is possible to perform a 'bootstrap' test of the stability of the output of an estimation algorithm, by running it multiple times on samples drawn with replacement from the original sample. However, while this can show that the output is stable, it does not show that the output is close to the truth, because the probability distribution might be unfaithful, or near-unfaithful, to the true causal graph. I recommend, as well, running search procedures on simulated data of the same size as the actual data, generated from a variety of initially plausible models. The results can give an indication of the probable accuracy of the search procedure and its sensitivity to search parameters and to the complexity of the data generating process. Of course, if the actual data is generated by a radically different structure, or if the actual underlying probability distribution or sampling characteristics do not agree with those in the simulations, these indications may be misleading. Also it should be kept in mind that even when a model suggested by an estimator fits the data very well, it is possible that there are other models that will also fit the data well and are equally compatible with background knowledge, particularly when the sample size is small. (These limitations are limitations on estimators of causal graphs or the effects of manipulations in general, and not just of the estimators described in this paper.)

What attitude should one have towards a causal model that is output by a search algorithm, and that passes all of these tests? Passing the tests is evidence in favor of the model, but there is no currently known method for quantifying how strong that evidence is. There is no known computationally feasible way of calculating the posterior probability of a causal model (although in simple cases without cycles, latent variables, or correlated errors it is possible to calculate the ratio of posterior probability of two causal models.) Since near-unfaithfulness is one important way in which the output at a given sample size can be far from the truth, one open question is whether it would be possible to quantify the degree of near-unfaithfulness of a population distribution to a causal graph, and then to calculate probabilistic bounds on the size of the error for a given maximum level of near-unfaithfulness. If it is possible to calculate probabilistic error bounds, then it would be possible to investigate the sensitivity of the probabilistic error bounds to variations in assumptions about maximum levels of near-unfaithfulness.

In many respects, the output of the causal model search algorithms is similar to the output of Gibbs sampling algorithms for estimating the expected value of a random variable. A Gibbs sampler is irreducible if it is possible to pass from any state with non-zero probability to any other state with non-zero probability in a finite number of transitions with non-zero probability. If a random variable has a finite expectation, and the Gibbs sampler is irreducible, it is guaranteed to converge to the expected value in the large sample limit with probability 1. However, proving that a Gibbs sampler is irreducible can be impossible in practice. Moreover, even if the irreducibility condition is met, at any given sample size it is not known how to put even probabilistic bounds on the size of the error (although lower bounds on the required sample size can sometimes be calculated.) If a Gibbs sampler is almost reducible, then the output of the Gibbs sampler can be very far off even at large sample sizes and even when the simulation gives every appearance of having approximately converged to a stable value (York 1992).

As an indication of how the PC algorithm performs on VAR models at realistic sample sizes, Demiralp and Hoover (2003) tested the PC algorithm on simulated data (which satisfied the assumptions needed for consistency) generated from a number of different structural VAR models at sample sizes of 500. They found that if the signal strength was large, the algorithm was very successful in its adjacency phase (approaching 100% as the signal strength grew), and somewhat less successful in the orientation phase (generally around 70% or more). These numbers are probably an upper limit to the performance of the algorithm on real data, because in real data the assumptions that guarantee the consistency of the algorithm may hold only approximately. Simulation tests of the PC algorithm for LSEMs are described in Spirtes, Glymour and Scheines (2000) with similar results. Correct background knowledge about temporal ordering can improve the reliability of the algorithms.

The problem of how to improve the performance of graph search algorithms on small samples, and how to extend searches to other kinds of graphical models is an area of active research. A variety of papers on this subject and applications of graph search algorithms can be found at <http://www2.sis.pitt.edu/~dsl/UAI/uai.html>. An implementation of the PC algorithm and other graphical model search algorithms can be found at <http://www.phil.cmu.edu/projects/tetrad/>. A modification of the PC algorithm is incorporated into the HUGIN software, available at <http://www.hugin.com/>. A number of different Bayesian network searches have been implemented by Kevin Murphy, and are available at <http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html>. Several useful introductory texts include Spirtes, Glymour and Scheines (2000), Pearl (2000), and Jordan (1999).

8 CONCLUSION

Interactions between the graphical causal modeling research program and the econometric causal modeling research program could potentially enrich both research programs. The research in the graphical causal modeling research program on automated model search and manipulation estimation, on theoretical questions about the limits of causal inference, and on what fundamental assumptions relating causation to probability should be made are applicable to econometric models. However, while the graphical causal modeling research program has achieved a great deal of generality in some areas of research with respect to distributional families, the research for the most part has been on a rather narrow range of kinds of causal interactions (those represented by DAGs, or occasionally DAGs with latent variables), The wide variety of econometric models and econometric causal inference methodology involving time series, equilibria, externally imposed constraints, systems deliberately chosen to minimize variance, contemporaneous causation, etc. could be an important source of methods, applications, and problems for the graphical causal modeling research program.

Peter Spirtes
Carnegie Mellon University
Ps7z@andrew.cmu.edu

ACKNOWLEDGEMENT

I would like to thank Kevin Hoover for valuable comments and suggestions.

NOTES

[1] In the general case, DAG models actually entail conditional independence relations among sets of variables. In the multi-variate normal case, all independence relations between sets of variables **X** and **Y** conditional on a set of variables **Z** are entailed by conditional independence relations among each individual variable $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ conditional on **Z**. This is not always the case for non-normal distributions.

[2] In counting degrees of freedom, it is assumed that no extra constraints (such as equality constraints among parameters) are imposed. The Bayesian Information Criterion for a DAG is defined as $\log P\left(\mathrm{D}\middle|\hat{\theta}_G,\mathrm{G}\right) - (\mathrm{d}/2)\log N$ where D is the sample data, G is a DAG, $\hat{\theta}_G$ is the vector of maximum likelihood estimates of the parameters for DAG G, N is the sample size, and d is the dimensionality of the model, which in DAGs without latent variables is simply the number of free parameters in the model.

[3] There are weaker versions of the Bayesian Causal Faithfulness Assumption (that assume a zero probability for zero partial correlations only between pairs of variables that are adjacent) that entail the existence of (weak) Bayes consistent estimators of the effects of manipulations, but at the cost of making the estimators more complicated and slower to compute. See Spirtes, Glymour and Scheines (2000), chapter 12.

[4] **X\Y** is a set whose members are the members of **X** that are not also members of **Y**.

REFERENCES 5

Bollen, K. (1989) *Structural Equations with Latent Variables*. New York: Wiley.

Bessler, David A. and Lee, Seongpyo (2002) 'Money and prices: US data 1869–1914 (a study with directed graphs)', *Empirical Economics* 27: 427–46.

Chickering, M. (2002) 'Optimal structure identification with greedy search', *Journal of Machine Learning Research* 3: 507–54.

Demiralp, S. and Hoover, K. (2003) 'Searching for the causal structure of a vector autoregression', *Oxford Bulletin of Economics and Statistics* 65: 745–67.

Hoover, K. (2001) *Causality in Macroeconomics*. Cambridge, UK: Cambridge University Press.

Hoover, K. (2003) ‘Non-stationary time series, co-integration, and the principle of the common cause’, *British Journal for the Philosophy of Science* 54: 527–51.

Jordan, M. (1999) *Learning Graphical Models*, 175–204, xx: MIT Press. 6

Kiiveri, H., Speed, T. and Carlin, J. (1984) ‘Recursive causal models’, *Journal of the Australian Mathematical Society* 36: 30–52.

Lauritzen, S. (2001) ‘Causal inference from graphical models’, in O. Barnsdorff-Nielsen, D. Cox and C. Kluppenlberg (eds) *Complex Stochastic Systems*, pp. 63–107, London: Chapman and Hall.

Lauritzen, S., Dawid, A., Larsen, B. and Leimer, H. (1990) ‘Independence properties of directed Markov fields’, *Networks* 20: 491–505.

Meek, C. (1995) ‘Causal inference and causal explanation with background knowledge’, in Besnard. Philippe and Hanks. Steve (eds) *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 403–10, San Mateo, CA: Morgan Kaufmann Publishers, Inc.

Moneta, A. (2003) *Graphical Models for Structural Vector Autoregressions*. xx: xxx. 7

Pearl, Judea (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Richardson, T. (1996) ‘A discovery algorithm for directed cyclic graphs’, in F. Jensen and E. Horvitz (eds) *Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Conference*, pp. 462–9, San Francisco: Morgan Kaufmann.

Robins, J., Scheines, R., Spirtes, P. and Wasserman, L. (2003) ‘Uniform consistency in causal inference’, *Biometrika* 90: 491–515.

Spirtes, P. (1995) ‘Directed cyclic graphical representation of feedback models’, in P. Besnard and S. Hanks (eds) *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann Publishers, Inc.

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction, and Search*, 2nd ed, Cambridge, MA: MIT Press.

Spirtes, P., Meek, C., Scheines, R. and Richardson, T. (1999) ‘An algorithm for causal inference in the presence of latent variables and selection bias’, in C. Glymour and G. Cooper (eds) *Computation, Causality, and Discovery*, pp. 211–52, xx: AAAI/MIT Press. 8

Strotz, R. and Wold, H. (1960) ‘Recursive versus nonrecursive systems: an attempt at synthesis’, *Econometrica* 28: 417–27.

Swanson, Norman R. and Granger, Clive W.J. (1997) ‘Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions’, *Journal of the American Statistical Association* 92: 357–67.

Verma, T. and Pearl, J. (1990) ‘Equivalence and synthesis of causal models’, in xxx (eds) *Proceedings of the Sixth Conference on Uncertainty in AI*. Mountain View, CA: Association for Uncertainty in AI, Inc. 9

Wright, S. (1934) ‘The method of path coefficients’, *Annals of Mathematical Statistics* 5: 161–215. 10

York, J. (1992) ‘Use of the Gibbs sampler in expert systems’, *Artificial Intelligence* 56: 115–30.

Yule, G. (1926) ‘Why do I sometimes get nonsensical relations between time-series? A study in sampling and the nature of time series’, *Journal of the Royal Statistical Society* 89: 1–64.

Authors Queries

Journal: **Journal of Economic Methodology**
Paper: **100210**
Title: **Graphical models, causal inference, and econometric models**

Dear Author
During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof. Many thanks for your assistance

Query Reference	Query	Remarks
1	Pearl (1988): Not listed in References.	
2	Figure 3: Please indicate where to place figure.	
3	The following are not listed in References: Andersson et al. (1995), Chickering (1995).	
4	Spirtes et al. 1993: Not listed in References.	
5	The following are not found in text: Bollen, K. (1989); Kiiveri, H., Speed, T. and Carlin, J. (1984).	
6	Jordan, M. (1999) Learning Graphical Models… xx: MIT Press: Please give city of publisher.	
7	Moneta, A., (2003) Graphical Models for Structural Vector Autoregressions, xx: xxxx: Please give publisher information.	

8	Spirtes, P., Meek, C., Scheines, R. and Richardson, T. (1999)… Discovery, pp. 211–52, xx: AAAI/MIT Press: Please give city of publisher.	
9	Verma, T. and Pearl, J. (1990)… in xxx (eds): Please give names of editors.	
10	Wright, S. (1934): Not found in text.	
11	Please supply figure 9.	