

2004

Causal Inference

Peter Spirtes
Carnegie Mellon University

Richard Scheines
Carnegie Mellon University

Clark Glymour
Carnegie Mellon University

Thomas Richardson
University of Washington

Christopher Meek
Carnegie Mellon University

Follow this and additional works at: <http://repository.cmu.edu/philosophy>

 Part of the [Philosophy Commons](#)

This Article is brought to you for free and open access by the Dietrich College of Humanities and Social Sciences at Research Showcase @ CMU. It has been accepted for inclusion in Department of Philosophy by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

Causal Inference

Peter Spirtes^{•¶}, Richard Scheines^{*}, Clark Glymour^{*¶}, Thomas Richardson[§], and Chris Meek[†]

1. Introduction

A principal aim of many sciences is to model causal systems well enough to provide insight into their structures and mechanisms, and to provide reliable predictions about the effects of policy interventions. In order to succeed in that aim, a model must be specified at least approximately correctly. Unfortunately, this is not an easy problem. Various kinds of information can be used to construct causal models: background knowledge about causal structures, statistical data gathered from experiments (randomized or non-randomized), and statistical data gathered from naturally occurring (random or non-random) samples. In order to transform the statistical data into causal conclusions it is necessary to have principles which link probability distribution to causal structures. In this article we will address the following questions:

1. What is the difference between a causal model and a statistical model? What uses do causal models have that statistical models do not have? (Section 2.)
2. What assumptions relating causality to probability should we make? (Section 3.)
3. What are the theoretical limits on causal inference under a variety of different assumptions about background knowledge? (Section 4.)
4. What are some sound methods of causal inference? (Section 5 when it is assumed that there are no latent causes, and section 6 when the possibility of latent causes is allowed.)
5. Are the methods of causal inference commonly employed in a variety of social sciences sound? (Section 7.)

Section 9 provides a guide to further reading about causal inference.

2. Causal Models and Statistical Models

2.1. *Conditioning and Manipulating*

There are two fundamentally different operations that transform probability distributions¹ into other probability distributions. The first is conditioning, which corresponds roughly to changing a probability distribution in response to finding out more information about the state of the world (or *seeing*). The second is manipulating, which corresponds roughly to changing a probability distribution in response to changing the state of the world in a specified way (or *doing*). (For more on the difference between conditioning and manipulating see Spirtes et al. 2000 and Pearl 2000). To illustrate the difference, we will consider a very simple example in which our pre-theoretic intuitions about causation are quite strong and uncontroversial. Consider a population of flashlights, each of which has

[•] Department of Philosophy, Carnegie Mellon University. [¶]Institute for Human and Machine Cognition. [§]Department of Statistics, University of Washington. [†]Microsoft Research.

¹ We are deliberately being ambiguous about the interpretation of “probability” here. The remarks here do not depend upon whether a frequentist, propensity, or personalist interpretation of “probability” is assumed.

working batteries and light bulbs, and a switch that turns the light on when the switch is *on* position, and turns the light off when the switch is in the *off* position. Each unit (flashlight) in the population has some properties (the switch position, and whether or not the light is on.) The properties are represented by the random variables *Switch* (*on* or *off*) and *Light* (*on* or *off*) respectively. The random variables have a joint distribution in the population. Suppose that in this case the joint distribution is the following:

$$P(\text{Switch}=\text{on},\text{Light}=\text{on}) = 1/2 \quad P(\text{Switch}=\text{off},\text{Light}=\text{off}) = 1/2 \\ P(\text{Switch}=\text{on},\text{Light}=\text{off}) = 0 \quad P(\text{Switch}=\text{off},\text{Light}=\text{on}) = 0$$

2.1.1. Conditioning

Conditioning is the more familiar operation, and much of statistics is devoted to finding efficient methods of estimating conditional probabilities. Given a randomly chosen flashlight, the probability that the bulb is *on* is 1/2. However, if someone observes that a flashlight has a switch in the *off* position, but does not directly observe whether the bulb is *off*, the probability of the light being *off* conditional on the switch being *off* is just the probability of the light being *off* in the subpopulation in which the switch is *off*, i.e. $P(\text{Light} = \text{off} | \text{Switch} = \text{off}) = P(\text{Light}=\text{off},\text{Switch}=\text{off})/P(\text{Switch}=\text{off}) = 1$. So conditioning transforms the joint distribution of the variables into a new probability distribution. Similarly, the probability of the switch being *off* conditional on the light being *off* is just the probability of the switch being *off* in the subpopulation in which the light is *off*, i.e. $P(\text{Switch} = \text{off} | \text{Light} = \text{off}) = P(\text{Light}=\text{off},\text{Switch}=\text{off})/P(\text{Light}=\text{off}) = 1$. An important feature of conditioning is that each conditional distribution is completely determined by the joint distribution (except when conditioning on sets of measure 0.)

2.1.2. Manipulating

Manipulating is also an operation that transforms joint probability distributions into other distributions. In contrast to conditioning, a *manipulated* probability distribution is not a distribution in a subpopulation of an existing population, but is a distribution in a (possibly hypothetical) population formed by externally *forcing* a value upon a variable in the system. Imagine that instead of observing that a switch was *off*, we set the switch to *off*. It follows that the probability of the light being *off* is 1. (Here, we are relying on pre-theoretic intuitions about the example to derive the correct values for the manipulated probabilities. In later sections, we will describe formal methods by which the manipulated probabilities can be calculated.) We will adapt the notation of Lauritzen (2001) and denote the post-manipulation probability of the light being *off* as $P(\text{Light}=\text{off} || \text{Switch}=\text{off})$, using a double-bar “||” for manipulation as distinguished from the single-bar “|” of conditioning.² Note that in this case, $P(\text{Light}=\text{off} || \text{Switch}=\text{off}) = P(\text{Light}=\text{off} | \text{Switch}=\text{off})$. Analogously to the notation for conditioning, one can also put a set of variable \mathbf{V} on the left side of the

² What time period after the manipulation does the “post-manipulation distribution” refer to? In this case, long enough for the system to reach an equilibrium. In cases where there is no equilibrium or some other time period is referred to, the relevant variables should be indexed by time explicitly.

manipulation double-bar, which represents the joint probability of \mathbf{V} after manipulating the variables on the right side of the manipulating double-bar.³

Suppose now that instead of manipulating the switch to *off*, we were to manipulate *Light* to *off*. Of course, the resulting probability distribution depends upon how we manipulated *Light* to *off*. If I manipulated *Light* to *off* by putting *Switch* to *off*, then of course the probability that *Switch* is *off* after the manipulation is equal to 1. On the other hand, if I were to manipulate *Light* to *off* by unscrewing the light bulb, the probability that *Switch* is *off* is 1/2, the same as the probability that it was *off* prior to my manipulation.

In the case where I manipulated *Light* to *off* by setting *Switch* to *off*, my manipulation directly affected one of the other variables (*Switch*) in the causal system. In the case where we manipulated *Light* to *off* by unscrewing the light bulb, our manipulation did not directly affect the other variables in the system (*Switch*). It is the latter sort of “ideal” manipulation that we will consider in the remainder of this article. This is not to indicate that “non-ideal” manipulations are not interesting, but simply that the best-developed theory is for “ideal” manipulations.

In the case where we perform an ideal manipulation of the light bulb to *off* (e.g. by unscrewing it), $1/2 = P(\mathbf{Switch}=\mathit{off} \parallel \mathit{Light}=\mathit{off}) \neq P(\mathbf{Switch}=\mathit{off} \parallel \mathit{Light}=\mathit{on}) = 1$. This illustrates two key features of manipulations.

The first is that in some cases the conditional probability is equal to the manipulated probability (e.g. $P(\mathit{Light}=\mathit{off} \parallel \mathbf{Switch}=\mathit{off}) = P(\mathit{Light}=\mathit{off} \parallel \mathbf{Switch}=\mathit{on})$) and in other cases, the conditional probability is not equal to the manipulated probability (e.g. $P(\mathbf{Switch}=\mathit{off} \parallel \mathit{Light}=\mathit{off}) \neq P(\mathbf{Switch}=\mathit{off} \parallel \mathit{Light}=\mathit{on})$). In this example, conditioning on *Light* = *off* raised the probability of *Switch* = *off*, but manipulating *Light* to *off* did not change the probability of *Switch* = *off*. In general, if conditioning on the value of a variable *X* raises the probability of a given event, manipulating *X* to the same value may raise, lower, or leave the same the probability of a given event. Similarly if conditioning on a given value of a variable lowers or leaves the probability of a given even the same, the corresponding manipulated probability may be higher, lower, or the same, depending upon the domain.

The second point is that even though *Light* = *on* if and only if *Switch* = *on* in the original population, the joint distributions that resulted from manipulating *Switch* and *Light* were different. It follows that, in contrast to conditioning, the results of manipulating depend upon more than the joint probability distribution. The “more than the joint probability distribution” that the results of a manipulation of a specified variable depend upon are causal relationships between variables. The reason that manipulating the switch position changed the status of the light is that the switch position is a cause of the status of the light; the reason that manipulating the light condition did not change the switch position is that the status of the light is not a cause of the switch position. Thus discovering (at least implicitly) the causal relations between variables is a necessary step to correctly inferring the results of manipulations.

³ We use boldface to represent sets of variables, and capitalized italics to represent variables and lower case italics to represent values of variables. There are a number of different alternative notations to the “||” in Spirtes et al. (2001) and Pearl (2001).

Conditional probabilities are typically of interest in those situations where the value of some variables (e.g. what bacteria are in your blood stream) are difficult to measure, but the values of other variables (e.g. what your temperature is, whether you have spots on your face) are easy to measure; in that case one can find out about the (probability distribution) of the value of the variable that is hard to measure by conditioning upon the values of the variables that are easy to measure.

Manipulated distributions are typically of interest in those situations where a decision is to be made, or a plan to be formulated. The possible actions that are considered in decision theory are typically manipulations, and hence the probability distributions that are relevant to the decision are manipulated probabilities, not conditional probabilities (although as we have seen, in some cases they may be equal.)

2.1.3. Other Kinds of Manipulations

Manipulating a variable to a particular value, e.g. $Switch = off$, is a special case of more general kinds of manipulations that can be performed. For example, instead of assigning a value to a variable, a probability distribution can be assigned to a variable X . This is what occurs in randomized experiments. Suppose that I randomize the probability distribution of $Switch$ to a distribution P' , where $P'(Switch = on) = 1/2$, and $P'(Switch = off) = 1/2$. In that case, we denote the manipulated probability of $Light$ as $P(Light \parallel P'(Switch))$, i.e. a probability distribution appears on the right hand side of the manipulation double bar. (The notation $P(Light=off \parallel Switch=off)$ is the special case where $P'(Switch=off) = 1$.)

More generally, given a set of variables \mathbf{V} , and manipulation of a set of variables $\mathbf{M} \subseteq \mathbf{V}$ to a distribution $P'(\mathbf{M})$, the joint distribution of \mathbf{V} after the manipulation is denoted $P(\mathbf{V} \parallel P'(\mathbf{M}))$. From $P(\mathbf{V} \parallel P'(\mathbf{M}))$ it is possible to form marginal distributions and conditional distributions among the variables in \mathbf{V} in the usual way.

In order to simplify the discussion, we will restrict our discussion of manipulation in several ways. First, we will not consider manipulations which assign a conditional probability distribution to a variable (e.g. $P'(Light=off \parallel Switch=on) = 1/2$, and $P'(Light=on \parallel Switch=on)=0$), rather than assigning a marginal distribution to that variable. Also, when multiple manipulations are performed, we will assume that in the manipulated distribution the variables are independent.

2.1.4. The Meaning of “Manipulation”

Note that while we have given examples of manipulations, we have not defined manipulation. “Manipulation” is one of a family of causal terms (including “direct cause”, “indirect cause”, “direct effect”, etc.) that are easily inter-definable, but not easily definable in terms of non-causal terms. A variety of different definitions of causal concepts in terms of non-causal concepts have been proposed, but they are typically both complicated and controversial. We will take a different approach here, and take the concept of “manipulation” as primitive, and introduce generally accepted axioms relating causal relations to probability distributions. The acceptability of these axioms does not depend upon the definition of causality. This will allow us to discuss principles of causal inference that are acceptable to a variety of schools of thought about the meaning of causality (just as there are at least some principles of probabilistic inference that do not depend upon the

definition of probability). Philosophical works about the meaning and nature of causation are Cartwright (1989 and 1999), Eells (1991), Hausman (1998), Shafer (1996), and Sosa and Tooley(1993).

2.2. Bayesian Networks – Causal and Statistical Interpretation

Bayesian networks are a kind of causal/statistical model that provide a convenient framework for representing and calculating the results of conditioning and manipulating. Bayesian networks also provide a convenient framework for discussing the relationship between causal relations and probability distributions. They are graphical models that generalize recursive structural equation models without correlated errors⁴, that have both a statistical and a causal interpretation. We will describe the statistical interpretation first, and then the causal interpretation, and finally the relationship between the two interpretations. Pearl (1988), Neapolitan (1990), Cowell (1999) and Jensen (2001) are introduction to Bayesian networks. Pearl (2001), Spirtes et al. (2001), and Lauritzen (2001) describe the relation between the causal and statistical interpretations.

2.2.1. Statistical Interpretation

A Bayesian network consists of two parts: a directed acyclic graph, and a set of parameters that map the graph onto a probability distribution via a rule that we will describe below. We will illustrate Bayesian networks using data from Sewell and Shah (1968) who studied five variables from a sample of 10,318 Wisconsin high school seniors. The variables and their values are:

<i>SEX</i>	[male = 0, female = 1]
<i>IQ</i> = Intelligence Quotient,	[lowest = 0, highest = 3]
<i>CP</i> = college plans	[yes = 0, no = 1]
<i>PE</i> = parental encouragement	[low = 0, high = 1]
<i>SES</i> = socioeconomic status	[lowest = 0, highest = 3]

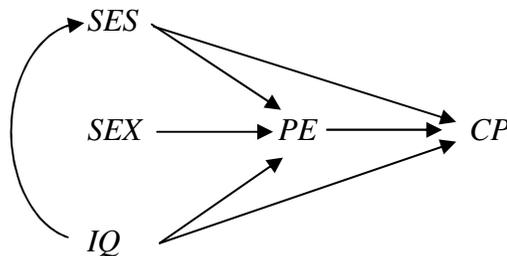


Figure 1: College Plans

The graph part of the Bayesian network that we will describe for the Sewell and Shah data is shown in Figure 1. We will explain the motivation behind hypothesizing the DAG in Figure 1 in section 5.

⁴ There are more general kinds of graphical models of which Bayesian networks are a special case that also have causal interpretations, but for the sake of simplicity, we postpone discussions of such models until later. See Whittaker (1990), Lauritzen (1996), and Edwards (2000), and Spirtes et al. (2000).

The following informal definitions describe various features of a directed graph.⁵ A directed graph consists of a set of vertices and a set of edges, where each edge is an ordered pair of vertices. In the example in Figure 1 the vertices are $\{IQ, SES, PE, CP, SEX\}$, and the edges are $\{IQ \rightarrow SES, IQ \rightarrow PE, SEX \rightarrow PE, SEX \rightarrow CP, SES \rightarrow PE, SES \rightarrow CP, PE \rightarrow CP\}$. In a directed graph, IQ is a **parent** of SES and SES is a child of IQ because there is an edge $IQ \rightarrow SES$ in the graph. $\mathbf{Parents}(G,V)$ denotes the set of parents of a vertex V in a directed graph G . A **directed path** in a directed graph is a sequence of edges all pointing in the same direction. For example, $IQ \rightarrow PE \rightarrow CP$ is a directed path from IQ to CP . Note that $SES \rightarrow PE \leftarrow SEX$ is not a directed path, because the two edges do not point in the same direction. CP is a **descendant** of IQ because there is a directed path from IQ to CP ; in addition, by convention, each vertex is a descendant 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.

A DAG over a set of variables \mathbf{V} **represents** any joint distribution that can be factored according to the following rule:

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V \mid \mathbf{Parents}(G,V)) \quad (1)$$

For example, the DAG in Figure 1 represents any joint probability distribution that can be factored according to the following formula:

$$P(SEX, IQ, SES, PE, CP) = P(IQ) \times P(SEX) \times P(SES \mid IQ) \times P(PE \mid SES, IQ, SEX) \times P(CP \mid PE, SES, SEX) \quad (2)$$

It is possible to associate a particular probability distribution with the DAG in Figure 1, by assigning values to the 80 free parameters. (Note that once $P(IQ=0)$ has been set, $P(IQ=1)$ is determined to be $1 - P(IQ=0)$ so it does not count as a free parameter. The same is true for the other possible combinations of variables values.) Once the values of these parameters are fixed, equation 1 can be used to compute the joint probability of any combination of values of the variables. For example,

$$P(SEX=0, IQ=1, SES=2, PE=1, CP=0) = P(IQ=1) \times P(SEX=0) \times P(SES=2 \mid IQ=1) \times P(PE=1 \mid SES=2, IQ=1, SEX=0) \times P(CP=0 \mid PE=1, SES=2, SEX=0).$$

Not every joint distribution $P(SEX, IQ, SES, PE, CP)$ can be expressed as a product of factors according to equation (2). Thus a Bayesian network that has the DAG shown in Figure 1 cannot represent an arbitrary probability distribution over $SEX, IQ, SES, PE,$ and CP .

By definition, a DAG G represents a probability distribution P if and only if P factors according to the DAG (equation (2).) The factorization of P according to G is equivalent to each variable X being independent of all the variables that are neither parents nor descendants of X in G , conditional on the parents of X in G . Let $I(\mathbf{X}, \mathbf{Y}, \mathbf{Z})_P$ mean that \mathbf{X} is independent of \mathbf{Y} conditional on \mathbf{Z} in distribution P , i.e. $P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})$. By convention, $I(\mathbf{X}, \mathbf{Y}, \emptyset)_P$ is trivially true. In the example of the DAG in Figure 1, for any

⁵ More formal definitions can be found in Spirtes et al. (2000), and Pearl (2000).

probability distribution that factors according to G , (i.e. satisfies equation (2)), the following conditional independence relations hold in P :

$$\begin{aligned} I(\{IQ\}, \emptyset, \{SEX\})_P & \quad I(\{SEX\}, \emptyset, \{IQ, SES\})_P & \quad I(\{SES\}, \{IQ\}, \{SEX\})_P & \quad (3) \\ I(\{PE\}, \{SES, IQ, SEX\}, \emptyset)_P & \quad I(\{CP\}, \{PE, SES, SEX\}, \{IQ\})_P \end{aligned}$$

These conditional independence relations hold regardless of what values are assigned to the free parameters associated with DAG G ; in that case we say that G **entails** the conditional independence relations. However, just because a conditional independence relation is not entailed by a DAG does not mean that it does not hold in any assignment of values to the free parameters: it just means that it does not hold in every assignment of values to the free parameters.

The conditional independence relations listed in (3) entail other conditional independence relations, e.g. $I(\{Vector\}, \{Algebra Skill, Real Analysis Skill\}, \{Statistics\})$. There is a graphical relationship, named d-separation, among three disjoint sets of vertices, which allows all of the conditional independence relations entailed by the Causal Markov Principle to be read off of the graph (Pearl, 1988, Lauritzen et al. 1990). The definition of d-separation is contained in the Appendix. For the purposes of this article, the important point is that there is a purely graphical relation “d-separation” such that if a DAG G represents a probability distribution $P(\mathbf{V})$, \mathbf{X} is d-separated from \mathbf{Y} conditional on \mathbf{Z} if and only if G entails that \mathbf{X} is independent of \mathbf{Y} conditional on \mathbf{Z} in $P(\mathbf{V})$.

2.2.2. Bayesian Networks - Causal Interpretation

The graph part of a Bayesian network also has a causal interpretation. Let a set of random variables \mathbf{S} be **causally sufficient** if \mathbf{S} does not omit variables, which are causes of any pair of variables in \mathbf{S} . Under the causal interpretation of a graph, a graph with causally sufficient set of variables \mathbf{S} **represents** the causal relations in a population N if there is a directed edge from A to B in the graph if and only if A is a direct cause of B relative to \mathbf{S} for some member of N . For example, under the causal interpretation of the graph in Figure 1 there is a directed edge from IQ to CP if and only if IQ is a direct (relative to the set of variables in the DAG) cause of CP for some member of the population.

The kind of causation that we are describing in this article is causation between variables (or kinds of events, e.g. *Switch* and *Light*) not between individual events (e.g. the event of a particular flashlight having a *Switch* value *on*, and the event of the same flashlight having the *Light* value *on*). Because the causal relation is between variables, and not between events, it is possible that each of two variables can cause each other. For example, pedaling a bicycle can cause the wheel to spin, and (on some kinds of bicycles) spinning the wheel can cause the pedal to move. Thus it is possible that a causal graph may be cyclic. The theory of cyclic causal graphs is important in such domains as econometrics, and is discussed in section 6.2.7.5, but it is also considerably more difficult and less developed than the theory of acyclic causal graphs. For the rest of this article we assume that all causal graphs are acyclic, unless we explicitly say otherwise.

A Bayesian network M has two parts: a DAG and a probability distribution that the DAG represents. We denote the DAG part of M as $G(M)$, and the distribution part of M as $P(M)$.

2.3. Structural Equation Models

Structural equation models are a kind of statistical/causal model widely used in the social sciences. There are several different methods of parameterizing structural equation models. The one described here is essentially the same as the one used in Bentler (1985). The variables in a linear structural equation model (SEM) can be divided into two sets, the “error variables” or “error terms,” and the substantive variables. The error terms are latent, and some of the substantive variables may be latent as well. A structural equation model consists of a set of structural equations, one for each substantive variable, and the distributions of the error terms; together these determine the joint distribution of the substantive variables. The structural equation for a substantive variable X_i is an equation with X_i on the left hand side of the equation, and the direct causes of X_i plus an error term ε_i on the right hand side of the equation. (In some representations, some substantive variables occur only on the right hand side of structural equations and do not have a corresponding error variable.) Bollen (1989) and Goldberger and Duncan (1973) are introductions to the theory of structural equation models.

Figure 2 contains an example of a latent variable SEM that is a model of mathematical test scores discussed in Spirtes et al. (2001). The original data set came from Mardia, Kent, and Bibby (1979). The test scores for 88 students in five subjects (*Mechanics*, *Vector Algebra*, *Algebra*, *Analysis* and *Statistics*) are the measured variables. The latent substantive variables are *Algebra Skill*, *Vector Algebra Skill*, and *Real Analysis Skill*. The distribution of the test scores is approximately normal. In Model M of Figure 2, the free parameters are the linear coefficient a , b , c , and d , and the variances and means of the error terms ε_M , ε_V , ε_{Ab} , ε_{An} , ε_S , δ_{Ab} , δ_{An} and δ_V . (The coefficients in the structural equations for *Mechanics*, *Algebra*, and *Statistics* have been fixed at 1 to ensure identifiability, which is explained below.) Note that in the equations we have used an assignment operator “:=” rather than an equals sign “=”. We do this to emphasize that the equations are structural, i.e. the quantity on the r.h.s. of the equation is not just equal to the l.h.s., but that it causes the l.h.s. Thus, as Lauritzen (2001) suggests, it is more appropriate to call these “structural assignment models” rather than “structural equation models”.

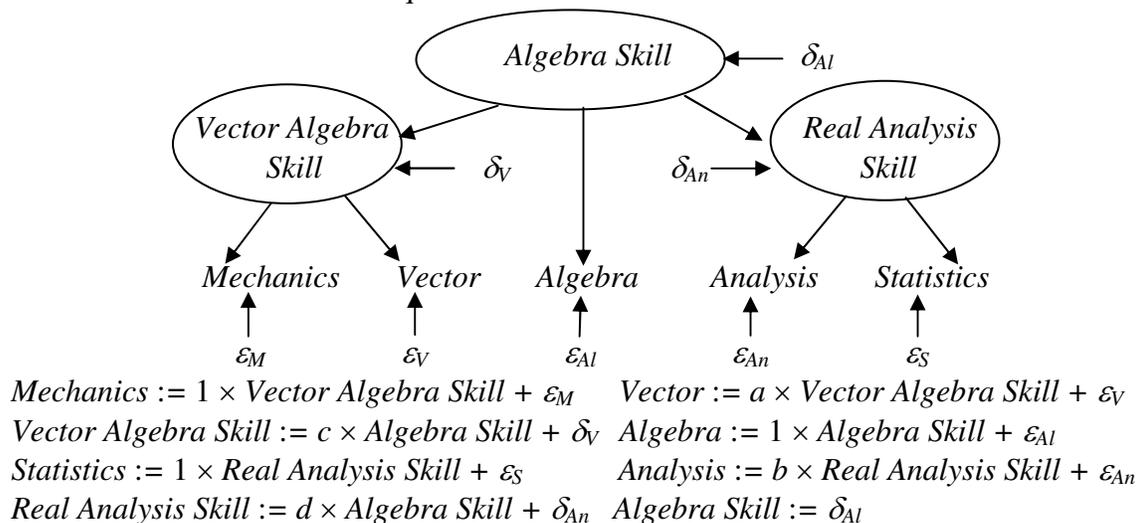


Figure 2: Mathematical Marks

Bayesian networks specify a joint distribution over variables with the aid of a DAG, while SEMs specify a value for each variable via equations. Superficially, they appear to be quite different. However, a SEM contains information about both the joint probability distribution over the substantive variables, and about the causal relations between the substantive variables. The joint distribution of the error terms, together with the equations, determines the joint distribution of the substantive variables. In addition, each SEM is associated with a graph (called a “path diagram”) that represents the causal structure of the model and the form of the equations, where there is a directed edge from X to Y ($X \rightarrow Y$) if Y is a (non-trivial) function of X (i.e. X is a direct cause of Y), and there is a bi-directed edge between the error terms ε_X and ε_Y if and only if the covariance between the error terms is non-zero. In path diagrams, latent substantive variables are enclosed in ovals. If the path diagram is acyclic, and there are no correlated errors, then a SEM is a special case of a Bayesian network, and it can be shown that the joint distribution factors according to equation (1), even when the equations are non-linear. Any probability distribution represented by the DAG in Figure 2, satisfies the following factorization condition, where f is the density:

$$f(\text{Mechanics}, \text{Vector}, \text{Algebra}, \text{Analysis}, \text{Statistics}) = f(\text{Mechanics} | \text{Vector Algebra Skill}) \times f(\text{Vector} | \text{Vector Algebra Skill}) \times f(\text{Algebra} | \text{Algebra Skill}) \times f(\text{Analysis} | \text{Real Analysis Skill}) \times f(\text{Statistics} | \text{Real Analysis Skill}) \times f(\text{Vector Algebra Skill} | \text{Algebra Skill}) \times f(\text{Real Analysis Skill} | \text{Algebra Skill}) \times f(\text{Algebra Skill}).$$

If the path diagram is cyclic or contains correlated errors, the factorization condition does not in general hold, but there are other properties of graphical models, which do still hold of SEMs, as, explained below. We denote the path diagram associated with a SEM M as $G(M)$, and the probability distribution associated with M as $P(M)$.

3. Causality and Probability

If any causal conclusions are to be drawn from the distribution of properties in a sample, some assumptions relating causal relations to probability distributions are required. In this section we will describe and discuss several such assumptions.

3.1. The Causal Markov Principle

We have described both a causal and statistical interpretation of graphs. What is the relationship between these two interpretations? We make the following assumption:

Causal Markov Principle (Factorization): For a causally sufficient set of variables \mathbf{V} in a population N , if an acyclic causal graph G represents the causal relations among \mathbf{V} in N , then G also represents $P(\mathbf{V})$, i.e.

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V | \text{Parents}(G, V)) \quad (1)$$

In the example of the causal DAG in Figure 1, the Causal Markov Principle implies

$$P(\text{SEX}, \text{IQ}, \text{SES}, \text{PE}, \text{CP}) = P(\text{IQ}) \times P(\text{SEX}) \times P(\text{SES} | \text{IQ}) \times P(\text{PE} | \text{SES}, \text{IQ}, \text{SEX}) \times P(\text{CP} | \text{PE}, \text{SES}, \text{SEX}) \quad (2)$$

An equivalent way of stating the Causal Markov Principle in terms of causal graphs is the following.

Causal Markov Principle (Independence): For a causally sufficient set of variables \mathbf{V} in a population N , if an acyclic causal graph G represents the causal relations among \mathbf{V} in N , each vertex X in \mathbf{S} is independent of the set of vertices that are neither parents nor descendants of X in G , conditional on the parents of X in G .

In the example of the causal DAG in Figure 1, the independence version of the Causal Markov Principle implies that the conditional independence relations listed in (3) hold in the probability distribution P in population N .

In a SEM with Gaussian error terms, \mathbf{X} is independent of \mathbf{Y} conditional on \mathbf{Z} if and only if for each $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, the partial correlation of X and Y given \mathbf{Z} (denoted $\rho(X, Y | \mathbf{Z})$) is equal to zero. The Causal Markov Principle is implicit in much of the practice of structural equation modeling (without cycles or correlated errors). For example, the only specific examples that Bollen (1989) gives of why a disturbance term for a variable X might be correlated with one of the causes of X other than sampling problems are all due to causal relations between the disturbance term and other causes of X . It is also often assumed that when correlated disturbance terms are introduced, this is because of some unspecified causal relation between the disturbance terms. Indeed, in the context of linear or non-linear structural equation models, the assumption that causally unconnected disturbance terms are independent entails the full Causal Markov Principle. Spirtes et al. (2001) contains a discussion of the Causal Markov Principle, and its limitations.

3.2. Calculating the Effects of Manipulations in Bayesian Networks

In a Bayesian network with causal DAG G , the effect of a manipulation can be calculated according to the following rule. If the distribution prior to the manipulation is $P(\mathbf{V})$, and the distribution after the manipulation is $P(\mathbf{V} \| P'(\mathbf{S}))$.

$$P(\mathbf{V} \| P'(\mathbf{S})) = P'(\mathbf{S}) \times \prod_{V \in \mathbf{V} \setminus \mathbf{S}} P(V | \mathbf{Parents}(G, V))$$

That is, into the original factorization of $P(\mathbf{V})$, one simply replaces

$$\prod_{S \in \mathbf{S}} P(S | \mathbf{Parents}(G, S))$$

by $P'(\mathbf{S})$, where \mathbf{S} is the set of manipulated variables. The manipulation operation depends upon what the causal graph is, because G appears in the term $P(S | \mathbf{Parents}(G, S))$. Also, because the value of S in the manipulation does not causally depend upon the values of the parents of S , the post-manipulation DAG that represents the causal structure does not contain any edges into S . (More general kinds of manipulations do not have this property.)

To return to the flashlight example, the pre-manipulation causal DAG is $Switch \rightarrow Light$, and the pre-manipulation distribution is:

$$\begin{aligned} P(Switch=on, Light=on) &= P(Switch=on) \times P(Light=on | Switch=on) = 1/2 \times 1 = 1/2 \\ P(Switch=off, Light=off) &= P(Switch=off) \times P(Light=off | Switch=off) = 1/2 \times 1 = 1/2 \\ P(Switch=off, Light=on) &= P(Switch=off) \times P(Light=on | Switch=off) = 1/2 \times 0 = 0 \end{aligned}$$

$$P(\text{Switch}=\text{on}, \text{Light}=\text{off}) = P(\text{Switch}=\text{on}) \times P(\text{Light}=\text{off} | \text{Switch}=\text{on}) = 1/2 \times 0 = 0$$

Suppose that *Light* is manipulated to the distribution $P'(\text{Light}=\text{off}) = 1$. Then the post manipulation distribution $P(\text{Switch}, \text{Light} | P'(\text{Light}=\text{off}) = 1)$ is found by substituting $P'(\text{Light})$ in for $P(\text{Light} | \text{Switch})$:

$$P(\text{Switch}=\text{on}, \text{Light}=\text{on}) = P(\text{Switch}=\text{on}) \times P'(\text{Light}=\text{on}) = 1/2 \times 0 = 0$$

$$P(\text{Switch}=\text{off}, \text{Light}=\text{off}) = P(\text{Switch}=\text{off}) \times P'(\text{Light}=\text{off}) = 1/2 \times 1 = 1/2$$

$$P(\text{Switch}=\text{off}, \text{Light}=\text{on}) = P(\text{Switch}=\text{off}) \times P'(\text{Light}=\text{on}) = 1/2 \times 0 = 0$$

$$P(\text{Switch}=\text{on}, \text{Light}=\text{off}) = P(\text{Switch}=\text{on}) \times P'(\text{Light}=\text{off}) = 1/2 \times 1 = 1/2$$

Although *Switch* and *Light* are symmetric in $P(\text{Light}, \text{Switch})$ the effects of manipulating them are asymmetric, because *Light* and *Switch* are not symmetric in the causal DAG. Manipulations in Bayesian networks are described in Spirtes et al. (2001), Pearl (2001), and Lauritzen (2001).

3.3. Manipulations in SEMs

In SEMs, there is a different, but equivalent representation of a manipulation. Suppose I were to manipulate the scores of all of the students by teaching them the answers to the questions on the *Analysis* test before they take it. The effect of this manipulation of the *Analysis* test score, on the joint distribution can be calculated by replacing the structural equation for *Analysis* with a new structural equation that represents its manipulated value (or more generally, the manipulated distribution of *Analysis*.) In this example, the structural equation $\text{Analysis} := b \times \text{Real Analysis Skill} + \varepsilon_{An}$ would be replaced by the equation $\text{Analysis} := 100$. The post-manipulation distribution is just the distribution entailed by the distribution of the error terms together with the new set of structural equations.

Of course, how an actual manipulation would affect other variables would depend upon how the system is manipulated. If the *Analysis* test score were changed by practice and studying to improve *Real Analysis Skill*, then one might expect that the *Statistics* test score might change as well. In that case, the manipulation has changed *Statistics*, but not through the mechanism of changing the *Analysis* score. The total effect of *Analysis* on *Statistics* is the resulting change in *Statistics* due to a kind of ideal manipulation of *Analysis*, in which the only variable in the system directly affected by the manipulation of *Analysis* is *Analysis* itself. Strotz and Wold suggested that an ideal manipulation of a variable X to a fixed value x could be represented by replacing the structural equation for X with a new equation $X = x$. The manipulated distribution is then just the distribution for the new structural equation model. The model that results from the manipulation of *Analysis* has a path diagram of the manipulated population formed by breaking all of the edges into the manipulated variable. In this case the edge from *Real Analysis Skill* to *Analysis* is removed, and the equation “ $\text{Analysis} := b \times \text{Real Analysis Skill} + \varepsilon_{An}$ ” is replaced by the equation “ $\text{Analysis} := 100$ ”.

In a linear SEM, the distinction is made between the direct effect of one variable on another and the total effect of one variable on another. The total effect of A on B measures how much B changes given a manipulation that makes a unit change in A . In the example of Figure 2, the total effect of *Algebra Skill* on *Vector* is given by $a \times c$, the product of the coefficient a associated with the edge from *Vector Algebra Skill* and the coefficient c

associated with the edge from *Algebra Skill* to *Vector Algebra Skill*. The direct effect of A on B is a measure of how much B changes given a manipulation that makes a unit change in A , while all variable other than A and B are manipulated to hold their current values fixed. The direct effect of A on B is given by the coefficient associated with the edge from A to B , or zero if there is no such edge. For example, the direct effect of *Vector Algebra Skill* on *Vector* is a , and the direct effect of *Algebra Skill* on *Vector* is zero.

In linear models, the total effect of A on B summarizes the effects manipulating A has on a variable B with a single number. In non-linear systems, such as Bayesian networks, this is generally not possible. The difference between $P(B)$ and $P(B|A)$ can depend both upon the value of B and the value of A .

Manipulations in SEMs are described in Strotz and Wold (1960), Spirtes et al. (2001), Pearl (2001), and Lauritzen (2001).

3.4. The Causal Faithfulness Principle

In this section, we motivate and describe further assumptions useful for making causal inferences. The Causal Faithfulness Prior Principle, the Causal Faithfulness Principle, and the Strong Causal Faithfulness Principle are three related assumptions of increasing strength. Which, if any, of these assumptions is accepted has strong consequences for the theoretical limits of inference of the effects of manipulations from measured data, as described in section 4.

The Causal Markov Principle states that causal structures entail conditional independence relations. If one is given a probability distribution, it is possible to use the Causal Markov Principle to conclude that certain causal relations exist. For example, *Switch* and *Light* are dependent. Hence the Causal Markov Principle entails that either *Switch* causes *Light*, *Light* causes *Switch*, or the system is not causally sufficient, i.e. there are unobserved variables that cause both *Light* and *Switch*. However, the Causal Markov Principle by itself does not entail that any variables are dependent. Hence observing independence between *Switch* and *Light* would not entail that *Switch* does not cause *Light* or *Light* does not cause *Switch*. It has been shown that in a SEM, given just the Causal Markov Principle, and allowing for the possibility of unmeasured common causes of measured variables, any direct effect of A on B is compatible with any covariance matrix among the measured variables (Robins et al. Forthcoming). Hence, in order to draw conclusions about direct effects from observational data, some additional assumptions must be made. We will examine three such assumptions in this section.

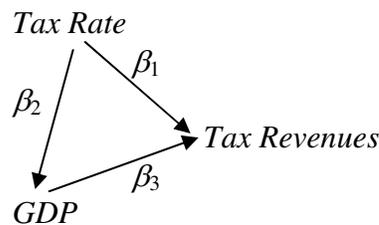


Figure 3: Distribution is Unfaithful to SEM when $\beta_1 = -(\beta_2\beta_3)$

If a probability distribution $P(\mathbf{V})$ is represented by a DAG G , then P is **faithful** to G if and only if every conditional independence relation in $P(\mathbf{V})$ is entailed by G (i.e. holds for all values of the free parameters, and not just some values of the free parameters.) Otherwise $P(\mathbf{V})$ is **unfaithful** to G . (In SEMs, this is equivalent to $P(\mathbf{V})$ is **faithful** to the path diagram of G if and only if every zero partial correlation that holds in $P(\mathbf{V})$ is entailed by G .) In the rest of the section, we will discuss faithfulness in SEMs, because the application to SEMs is somewhat simpler than the application to Bayesian networks.

Figure 3 shows an example of how an unfaithful distribution can arise. For example, suppose the DAG in Figure 3 represents the causal relations among the *Tax Rate*, the *Economy*, and *Tax Revenues*. In this case there are no vanishing partial correlation constraints entailed for all values of the free parameters. But if $\beta_1 = -(\beta_2 \times \beta_3)$, then *Tax Rate* and *Tax Revenues* are uncorrelated, even though the path diagram does not entail that they are uncorrelated. The SEM postulates a direct effect of *Tax Rate* on *Revenue* (β_1), and an indirect effect through the *Economy* ($\beta_2 \times \beta_3$). The parameter constraint indicates that these effects *exactly* offset each other, leaving no total effect whatsoever.

It is clear from this example that unfaithful distributions greatly complicate causal inference. Although, *Tax Rate* and *Tax Revenues* are completely uncorrelated, there can be a strong direct and total effect of *Tax Rate* on *Tax Revenues*. This would be particularly difficult to detect if *GDP* were unmeasured. The violation of faithfulness described in the example only occurs for very special values of the parameters, i.e. $\beta_1 = -(\beta_2 \times \beta_3)$. In general, the set of free parameter values for any DAG that lead to unfaithful distribution is zero, for any “smooth” prior probability distribution⁶ (e.g. Normal, exponential, etc.) over the free parameters. This motivates the following Bayesian assumption. (The methods for, and consequences of, assigning prior probabilities to causal graphs and parameters in order to perform Bayesian inferences are described in more detail in section 4. Although we state these assumptions for SEMs for convenience, there are more general versions of these assumptions, which apply to Bayesian networks in general.)

Causal Faithfulness Prior Principle: Suppose that there is a population N with distribution $P(\mathbf{V})$, and a DAG G that represents the causal relations in N . The set of parameters for which there is no edge between a pair of variables X and Y , but for which there exists a set \mathbf{Z} not containing X or Y such that $\rho(X, Y | \mathbf{Z}) = 0$, has prior probability zero.

This assumption is implicitly made by any Bayesian who has a prior over the parameters taken from the usual families of distributions. Of course, this argument is not relevant to those who reject Bayesian arguments, nor to Bayesians whom place a prior over the parameters that is not “smooth”, and assign a non-zero probability to violations of Faithfulness. And there are cases where it would not be reasonable to assume the Causal Faithfulness Principle. For example, in addition to “cancellations”, deterministic relationships among variables can lead to violations of the Causal Faithfulness Principle.

⁶ A “smooth” prior is absolutely continuous with Lebesgue measure.

A stronger version of the Causal Faithfulness Prior Principle, which does not require acceptance of the existence of prior probability distributions is the following.⁷

Causal Faithfulness Principle (SEMs): Suppose that there is a population N with distribution $P(\mathbf{V})$, and a DAG G that represents the causal relations in N . If for any set \mathbf{Z} not containing X or Y , $\rho(X, Y|\mathbf{Z}) = 0$, then the linear coefficient of X in the equation for Y is zero (i.e. there is no edge from X to Y , and X is not a cause of Y).

The Causal Faithfulness principle is a kind of simplicity assumption. If a distribution P is faithful to a SEM M_1 without latent variables or correlated errors, and P also results from assigning parameter values to another SEM M_2 to which P is not faithful, then M_1 has fewer free parameters than M_2 .

The Faithfulness principle limits the SEMs considered to those in which population constraints are entailed by structure, not by particular values of the parameters. Faithfulness should not be assumed when there are deterministic relationships among the substantive variables, or equality constraints upon free parameters, since either of these can lead to violations of the assumption. Some form of the assumption of Faithfulness is used in every science, and amounts to no more than the belief that an improbable and unstable cancellation of parameters does not hide real causal influences. When a theory cannot explain an empirical regularity save by invoking a special parameterization, most scientists are uneasy with the theory and look for an alternative.

The Causal Faithfulness Principle, as stated, has implications only for cases where a partial correlation is exactly zero. It is compatible with a partial correlation being arbitrarily small, while an edge coefficient is arbitrarily large. A stronger version of the Causal Faithfulness Principle eliminates this latter possibility.

Strong Causal Faithfulness Principle (SEMs): Suppose that there is a population N with distribution $P(\mathbf{V})$, and a DAG G that represents the causal relations in N . If for any set \mathbf{Z} not containing X or Y , $\rho(X, Y|\mathbf{Z})$ is small, then the linear coefficient of X in the equation for Y is small.

This is not a precise statement but could be made precise in several ways. One way in which it could be made precise is to assume that the linear coefficient of X in the equation for Y is no more than some constant k times $\rho(X, Y|\mathbf{Z})$, for any \mathbf{Z} .

Unlike the Causal Faithfulness Principle, violations of the Strong Causal Faithfulness Principle are not probability zero for every “smooth” prior over the parameters. However, the Strong Causal Faithfulness Principle is implicit in common modeling practices. For example, it is often the case that in causal modeling in various fields, a large number of measured variables \mathbf{V} is reduced by regressing some variable of interest Y on the other variables, and eliminating from consideration those variables that have small regression coefficients. Since a small regression coefficient of Y when X is regressed on $\mathbf{V}\setminus\{X\}$ entails

⁷ This is a stronger assumption, because it eliminates all parameters that lead to violations of faithfulness from the sample space, instead of simply leaving them in the sample space and assigning them prior probability zero.

that $\rho(X, Y | \mathbf{V} \setminus \{X, Y\})$ is small, this amounts to assuming that a small partial correlation entails a small effect of Y on X .

The various forms of the Causal Faithfulness Principle are described and discussed in Spirtes et al. (2000).

4. Causal Inference

In all the examples we discuss, the Causal Markov Principle is assumed. We also assume that there is a causally sufficient set of variables \mathbf{V} that is joint Normal, or contains all discrete variables. We will assume that the causal graph is acyclic unless explicitly stated otherwise. We also assume that there are no correlated errors, unless explicitly stated otherwise (this case will be discussed further in the section on latent variables.)

4.1. Point Estimation

4.1.1. The Classical Framework

In the classical framework, an estimator $\hat{\theta}_n$ is a function that maps samples of size n into real numbers. An estimator is a *pointwise consistent* estimator of a causal parameter θ if, for each possible value of θ , in the limit as the sample size approaches infinity, the probability of the distance between the estimator and the true value of θ being greater than any fixed finite value approaches zero. However, pointwise consistency is only a guarantee about what happens in the limit, not at any finite sample size. Pointwise consistency is compatible with there being at each sample size, some value of the causal parameter such that the probability of the estimator being far from the true value is high. Suppose that one were interested in answering questions of the following kind: What sample size is needed to guarantee that, regardless of the true value of the causal quantity, it is “improbable” that the estimator is “far” from the truth? “Improbable” and “far” are vague terms, but they can be made precise. “Improbable” can be made precise by choosing a positive real ε , such that any probability less than ε is improbable. “Far” can be made precise by choosing a positive real δ such that any distance greater than δ is “far”. Then the question can be rephrased as follows: What sample size is needed to guarantee that regardless of the true value of the causal quantity, the highest probability that my estimator was more than δ away from the truth was less than ε ? Given only pointwise consistency, the answer may be “infinite”. However, a stronger form of consistency, **uniform consistency**, guarantees that answers to questions of the form given above are always finite, for any given ε and δ . See Bickel and Doksum (2001).

Both pointwise and uniformly consistent estimators are asymptotically unbiased (i.e. as n approaches infinity, the difference between the expected value of $\hat{\theta}_n$ and θ is zero.) Another desirable feature of an estimator is that it have low variance (because the mean squared error of the estimator is equal to the square of the bias plus the variance of the estimator.)

4.1.2. The Bayesian Framework

In the Bayesian framework, point estimation of a manipulation proceeds by:

1. Assigning a prior probability to each causal graph.
2. Assigning joint prior probabilities to the parameters conditional on a given causal graph.
3. Calculating the posterior probability of the manipulation (which is a function of the posterior probabilities of the graphs and the graph parameters.)
4. Turning the posterior probability over the manipulation into a point estimate by e.g. returning the mean or median value of the estimate.

Note that such a Bayes estimator is a function not only of the data, but also of the prior probabilities. Such a Bayes estimator can have a weaker sense of consistency than pointwise consistency. If the set of causal hypotheses (graph – probability distribution pairs) for which the Bayes estimator converges to the correct value has a posterior probability of 1, then we will say that it is Bayes consistent (with respect to the given set of priors.) Since a pointwise consistent estimator converges to the correct value for all causal hypotheses in the sample space, pointwise consistency entails Bayes consistency. The Bayesian approach to causal inference is described in Heckerman (1998).

5. No Unmeasured Common Causes

First, we consider the case where there is a causally sufficient set of variables \mathbf{V} that are all measured.

5.1. *Known Causal Graph*

There are uniformly consistent estimators of the parameters for multivariate Normal or discrete DAG models. In the case of multivariate Normal distributions, a maximum likelihood estimate of the edge coefficients can be obtained by regressing each variable on its parents in the causal DAG. In the case of discrete DAG models, a maximum likelihood estimate of the parameters $P(V|\mathbf{Parents}(G,V))$ can be obtained by using the relative frequency of V conditional on $\mathbf{Parents}(G,V)$.

As we saw in sections 3.2 and 3.3 it was shown that $P(\mathbf{V}||P'(V))$ is a rational function of the parameters. Hence, it follows that there are uniformly consistent estimates of $P(\mathbf{V}||P'(V))$. (We assume in the case of discrete variables that $P'(V)$ does not assign a non-zero probability to any value of V that has probability 0 in the unmanipulated population.) ???

5.2. *Unknown Causal Graph*

$P(\mathbf{V}||P'(V))$ in a population is a function of the causal graph for that population. If the causal graph were in turn a function of the population $P(\mathbf{V})$, then $P(\mathbf{V}||P'(V))$ could still be calculated from $P(\mathbf{V})$ alone. We will examine under what conditions the causal graph is a function of $P(\mathbf{V})$. We will also determine what inferences can be made when it is possible to infer some, but not all features of the causal graph from $P(\mathbf{V})$.

5.2.1. **The Causal Faithfulness Principle**

If the Causal Markov Principle is assumed, but neither the Causal Faithfulness Principle, nor the Causal Faithfulness Principle is assumed, then the orientation of edges is completely undetermined. This is because any (multivariate Normal or discrete)

distribution can be represented by some submodel of a DAG in which any two vertices are adjacent, regardless of the orientation of the edges. It follows that if X and Y are correlated, then it is *never* the case that there is a consistent estimator of either $P(Y|P'(X))$ or $P(X|P'(Y))$ that is a function of $P(\mathbf{V})$ (although interval estimates may be possible.) See Robins et al. (forthcoming), Spirtes et al. (2001), and Zhang (2002).

However, if any of the Causal Faithfulness Principles is assumed, in many cases some orientations of edges are incompatible with $P(\mathbf{V})$, and considerably more information about the effects of a manipulation can be derived from $P(\mathbf{V})$, as described in the subsequent sections.

5.2.2. Distribution Equivalence

Consider the College Plans example. There are a variety of ways of scoring such a discrete model, which include $p(\chi^2)$, and the BIC or Bayes Information Criterion (see section 5.2.7, and Bollen and Long 1993.) In order to evaluate how well the data supports this *causal* model, it is necessary to know whether or not there are other *causal* models compatible with background knowledge that fit the data equally well. In this case, for each of the path diagrams in Figure 4, and for *any* data set D , the two models fit the data equally well, and receive the same score. If \mathbf{O} represent the set of measured variables in path diagrams G_1 and G_2 , then G_1 and G_2 are **distribution equivalent over \mathbf{O}** if and only if for every model M such that $G(M) = G_1$, there is a model M' with graph $G(M') = G_2$, and the marginal of $P(M)$ over \mathbf{O} equals the marginal of $P'(M')$ over \mathbf{O} , and vice-versa.⁸ (Informally, any probability distribution over \mathbf{O} generated by an assignment of values to free parameters of graph G_1 can be generated by a assigning values to the free parameters of graph G_2 , and vice-versa.) If G_1 and G_2 have no latent variables then we will simply say that G_1 and G_2 are **distribution equivalent**. If two distribution equivalent models are equally compatible with background knowledge, and have the same degrees of freedom the data does not help distinguish them, so it is important to be able to find the complete set of path diagrams that are distribution equivalent to a given path diagram. (Each model that contains a path diagram in has the same number of degrees of freedom.)

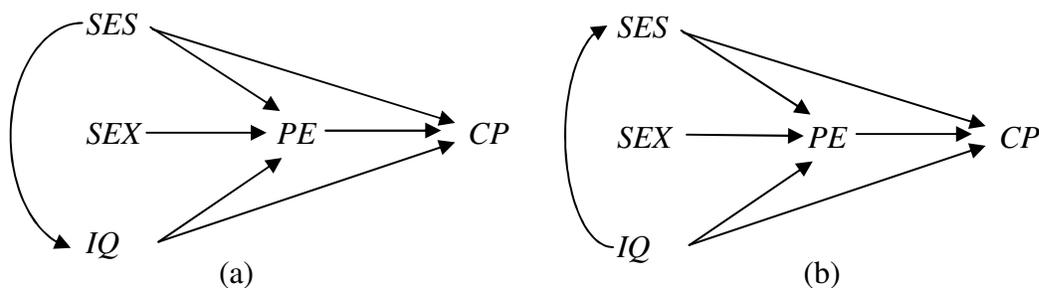


Figure 4: Simple Distribution Equivalence Class

⁸ For technical reasons, a more formal definition requires a slight complication. G is a **subgraph** of G' when G and G' have the same vertices, and G has a subset of the edges in G' . G_1 and G_2 are **distribution equivalent over \mathbf{O}** if for every model M such that $G(M) = G_1$, there is a model M' with path diagram $G(M')$ that is a sub-path diagram of G_2 , and the marginal over \mathbf{O} of $P(M')$ equals the marginal over \mathbf{O} of $P(M)$, and for every model M such that $G(M') = G_2$, there is a model M with graph $G(M)$ that is a subgraph of G_1 , and the marginal over \mathbf{O} of $P(M)$ equals the marginal over \mathbf{O} of $P(M')$.

As we will illustrate below, it is often far from obvious what constitutes a complete set of path diagrams distribution equivalent to a given graph, particularly when there are latent variables, cycles, or correlated errors. We will call such a complete set a **distribution equivalence class over \mathbf{O}** . (Again, if we consider only models without latent variables, we will call such a complete set a **distribution equivalence class**.) If it is a complete set of graphs without correlated errors or directed cycles, i.e. directed acyclic graphs, that are distribution equivalent we will call it a **simple distribution equivalence class over \mathbf{O}** .

5.2.3. Features Common to a Simple Distribution Equivalence Class

An important question that arises with respect to simple distribution equivalence classes is whether it is possible to extract the features that the set of simple distribution equivalent path diagrams have in common. For example, each of the graphs in Figure 4 has the same adjacencies. The edge between IQ and SES points in different directions in the two graphs in Figure 4. However, $PE \rightarrow CP$ is the same in both members of the simple distribution equivalence class. This is informative because even though the data does not help choose between members of the simple distribution equivalence class, insofar as the data is evidence for the disjunction of the members in the simple distribution equivalence class, it is evidence for the orientation $PE \rightarrow CP$.

In section 5.2.5 we describe how to extract all of the features common to a simple distribution equivalence class of path diagrams.

5.2.4. Distribution Equivalence for Path diagrams Without Correlated Errors or Directed Cycles

If for causal model M , there is another causal model M' with a different causal graph but the same number of degrees of freedom, and the same marginal distribution over the measured variables in M , then the $p(\chi^2)$ for M' equals $p(\chi^2)$ for M , and they have the same BIC scores. Such models are guaranteed to exist if there are models that have the same number of degrees of freedom and contain graphs that are distribution equivalent to each other. Theorem 1 (Verma and Pearl 1990, Spirtes et al. 2000) shows how distribution equivalence can be calculated in $O(E^2)$ time, where E is the number of edges in a path diagram. X is an **unshielded collider** in directed acyclic graph G if and only if G contains edges $A \rightarrow X \leftarrow B$, and A is not adjacent to B in G .

Theorem 1: For multivariate normal distributions, or discrete distributions, two causal models with acyclic causal graphs but no correlated errors or cycles are distribution equivalent if and only if they contain the same vertices, the same adjacencies, and the same unshielded colliders.

5.2.5. Extracting Features Common to a Simple Distribution Equivalence Class

Theorem 1 is also the basis of a representation (called a pattern in Verma and Pearl 1990) of the entire set of graphs without correlated errors or cycles distribution equivalent to a given graph without correlated errors or cycles. The pattern for each DAG in Figure 4 is shown in Figure 5.

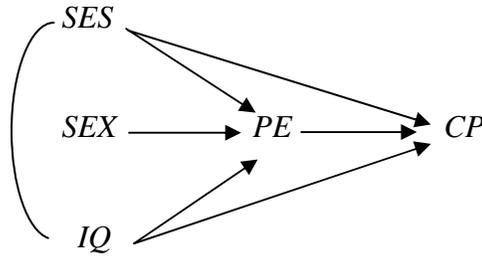


Figure 5: Pattern

A pattern has the same adjacencies as the DAGs in the simple distribution equivalence class that it represents. In addition, an edge is oriented as $X \rightarrow Z$ in the pattern if and only if it is oriented as $X \rightarrow Z$ in every DAG in the simple distribution equivalence class. Meek (1995), Andersson et al. (1995), and Chickering (1995) show how to generate a pattern from an acyclic graph in $O(E)$ time (where E is the number of edges.) Section 5.2.7.4 discusses the problem of constructing a pattern from sample data.

5.2.6. Calculating the Effects of Manipulations from a Pattern

The rules that specify which effects of manipulations can be calculated from a pattern and how to calculate them, and which effects of manipulations cannot be calculated from a pattern are described in Spirtes et al. (2000). Here we give some examples without proof.

Suppose that it is known that the pattern in Figure 5 is the true causal pattern (i.e. the true causal DAG is represented by that pattern.) The pattern represents the two DAGs in Figure 4. The DAG in Figure 4(a) predicts that $P(IQ \parallel P'(SES)) = P(IQ)$, because IQ is not an effect of SES . However, the DAG in Figure 4(b) predicts that $P(IQ \parallel P'(SES)) = P(IQ)$, because IQ is an effect of SES in Figure 4(b). Hence, knowing only that the true causal DAG is represented by the pattern in Figure 5 does not determine a unique answer for the value of $P(IQ \parallel P'(SES))$, and there are no consistent estimators of $P(IQ \parallel P'(SES))$ that are functions of just the joint distribution of $P(IQ, SEX, SES, PE, CP)$ and the pattern in Figure 5.

In contrast, both of the DAGs in Figure 4 predict that $P(PE \mid SES, IQ \parallel P'(IQ)) = P(PE \mid SES, IQ)$ (where $P(PE \mid SES, IQ \parallel P'(IQ))$ denotes the distribution of PE conditional on SES and IQ , after IQ has been manipulated to $P'(IQ)$.) Hence there are uniformly consistent estimators of $P(PE \mid SES, IQ \parallel P'(IQ))$ that are functions of just the joint distribution of $P(IQ, SEX, SES, PE, CP)$ and the pattern in Figure 5.

Finally, there are conditional distributions which do change under manipulation, but that can be calculated from quantities that do not change under manipulation. For example,

$$P(CP \mid PE \parallel P'(PE)) = \sum_{SEX, SES} P(CP \mid PE, SES, IQ) \times P(IQ \mid SES) \times P(SES) \times P'(PE)$$

Given the Sewell and Shah data, and assuming that the pattern in Figure 5 is the correct pattern, the following are estimates of $P(CP \mid PE \parallel P'(PE))$:

$$\begin{aligned} P(CP=0 \mid PE=0 \parallel P'(PE)) &= .095 & P(CP=1 \mid PE=0 \parallel P'(PE)) &= .905 \\ P(CP=0 \mid PE=1 \parallel P'(PE)) &= .484 & P(CP=1 \mid PE=1 \parallel P'(PE)) &= .516 \end{aligned}$$

5.2.6.1. Bayes, Pointwise, and Uniformly Consistent Estimators

Suppose that the only given data are samples from a jointly Normal probability distribution, and that the set of variables is known to be causally sufficient. Under what assumptions and conditions are there Bayes, pointwise, or uniformly consistent estimators of manipulations that are functions of the sample (where the causal DAG is unknown)?

If E is an estimator of some quantity Q then under their standard definitions, Bayes, pointwise and uniform consistency of E require that as the sample size increases E approaches Q , regardless of the true value of Q . Under this definition there are no consistent estimators of any effects of any manipulation even given the Strong Causal Faithfulness principles. However, given either the Causal Faithfulness Prior Principle, the Causal Faithfulness Principle, or the Causal Faithfulness or Strong Causal Faithfulness Principles, there are slightly weakened senses of Bayes, pointwise and uniform consistency respectively under which there are consistent estimators of the effects of some manipulations. In the weakened sense, an estimator can return “don’t know” as well as a numerical estimate, and a “don’t know” estimate is considered to be zero distance from the truth. In order for an estimator to be non-trivial, there must be some values of Q for which, with probability 1, in the large sample limit the estimator does not return “don’t know”. From now on, we will use “Bayes consistent estimator”, “pointwise consistent estimator” and “uniformly consistent estimator” in this weakened sense.

Suppose that we are given a causally sufficient set of Normally distributed or discrete variables \mathbf{V} and the Causal Markov Principle, but not any version of the Causal Faithfulness Principle. If the time order is known, and there are no deterministic relations among the variables, then there are uniformly consistent estimators of any manipulation. If the time order is not known, then for any X and Y that are dependent, regardless of what the true probability distribution $P(\mathbf{V})$ is, there are no Bayes, pointwise or uniformly consistent estimators of $P(Y|P'(X))$. This is because there are always a DAG compatible with the Causal Markov Principle in which X is a cause of Y , and another DAG in which X is not a cause of Y .

The following table summarizes the results about what kinds of consistent estimators exist under which assumptions. In all cases it is assumed that the Causal Markov Principle is true, there are no deterministic relations among variables, and all distributions are multivariate normal or all variables are discrete. Some combinations of conditions are missing because the Strong Causal Faithfulness Principle entails the Causal Faithfulness Principle, which entails the Causal Faithfulness Prior Principle. The first 4 columns are combinations of assumptions that are possible, and the last three columns give the consequences of those principles. Not surprisingly, the stronger the version of Causal Faithfulness that is assumed, the stronger the sense of consistency that can be achieved.

Time order	Causal Faithfulness Prior	Causal Faithfulness	Strong Causal Faithfulness	Existence of Bayes Consistent	Existence of Pointwise Consistent	Existence of Uniformly Consistent
NO	NO	NO	NO	NO	NO	NO
NO	YES	NO	NO	YES	NO	NO
NO	YES	YES	NO	YES	YES	NO

NO	YES	YES	YES	YES	YES	YES
YES	NO	NO	NO	YES	YES	YES

Table 1

We will describe the construction of consistent estimators of manipulations in 5.2.7. Even given the Strong Causal Faithfulness Principle, because all of the DAGs represented by a given pattern are distribution equivalent, the most that can be determined even in the large sample limit from the data is which pattern is correct. So any consistent estimator is sometimes going to return “don’t know.” In general, consistent estimators return numerical estimates (as opposed to “don’t know”) whenever the value of the manipulation is a function of the true pattern and the true distribution (as described in section 5.2.6). The results in Table 1 are proved in Robins et al. (forthcoming), Spirtes et al. (2001), and Zhang (2002).

5.2.7. Model Selection and Construction

In this section, we discuss some of the methodological implications of the results presented in the previous sections for model construction. The proper methodology depends upon what the model’s intended use is: calculating conditional probabilities, or calculating manipulations or the true causal pattern. Throughout we will use the College Plans data as an example. Analogous methodological conclusions can be drawn for SEMs. At this point we will not consider the issues of how the variables in the College Plans data set were constructed, nor will we impose any constraints on the models drawn from background knowledge. Such further considerations could be incorporated into the search algorithms discussed below, which could alter their output.

5.2.7.1. Estimation of Conditional Probabilities

Suppose that a model of the College Plans data is to be used to predict the value of CP from the other observed variables. One way to do this is to estimate $P(CP|SEX,IQ,SES,PE)$ and choose the value of CP with the highest probability. The relative frequency of CP conditional on SEX , IQ , SES , and PE is a uniformly consistent estimator of $P(CP|SEX,IQ,SES,PE)$. If the sample size is large, then the relative frequency will be a good estimator of $P(CP|SEX,IQ,SES,PE)$; however, if the sample size is small, then it typically will not be a good estimator because the number of sample points with fixed given values for SEX , IQ , SES , and PE will be small or possibly zero, and the estimator will have very high variance and a high mean squared error. (If the variables were continuous, the analogous operation would be to regress CP on SEX , IQ , SES , and PE .) In that case, a number of machine learning techniques, including variable selection algorithms, neural networks, support vector machines, decision trees, and non-linear regression, can be applied. (See Mitchell 1997). Once the model is constructed it can be evaluated in several different ways. For example, the sample can be divided into a training set and a test set. Then the model can be constructed on a training set, and then the mean squared error can be calculated on the test set. There are also a variety of other cross-validation techniques that can be used to evaluate models. If several different models are constructed, the one with the smallest mean squared error on the test set can be chosen. Note that it does not matter if there are several different models that predict CP equally well: in that case, any of them can be used. If the goal is to predict CP from PE alone, then the sample size is large

enough that the relative frequency of CP conditional on PE is a good estimator of $P(CP|PE)$.

On the other hand, if the model is to be used to understand the processes by which CP are determined, or to predict the effect on CP of a manipulation of PE (i.e. $P(CP|P'(PE))$), then a very different methodology is called for. In general, the relative frequency of CP conditional on PE is not a uniformly or pointwise consistent estimator of $P(CP|P'(PE))$, even if PE is known to precede CP . As pointed out in section 5.2.3, however, there are Bayes, pointwise or uniformly consistent estimators (depending upon whether one assumes the Causal Faithfulness Prior, the Causal Faithfulness, or the Strong Causal Faithfulness Principle.) Such consistent estimators can be constructed in a two-step process. First, perform a search that in the large sample limit returns the correct pattern. Second, if the effect of a manipulation is to be calculated, calculate it from the pattern and the sample distribution (as described in section 5.2.6.) Here we will describe searches that return the correct pattern in the large sample limit.

Even if latent variables are excluded, the search space is enormous (the number of different models grows super-exponentially with the number of variables.) Of course background knowledge, such as time order, can vastly reduce the search space. Nevertheless, even given background knowledge, the number of a priori plausible alternatives is often orders of magnitude too large to search by hand.

The problem is made even more difficult because there is in practice no simple way to test how well the estimation procedure has performed on a given sample, even under the Strong Causal Faithfulness Principle. The procedures are so complex that an analytic calculation of error is computationally infeasible. 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.

5.2.7.2. *Bayesian Causal Inference*

One kind of solution to the problem of finding causal structure, or calculating the effects of a manipulation is a Bayesian solution. In the Bayesian framework a prior probability is assigned to the space of hypotheses, and then a posterior probability on the space of alternative is calculated. The space of alternatives can either be the space of causal DAGs (and their associated parameters), or the space of patterns (and their associated parameters).

Under the family of priors described in Heckerman (1998), with probability 1 the posterior of the true causal DAG will not be smaller than the posterior of any other DAG. However, if the space of alternatives is causal DAGs, then in general the posterior probability will not converge to 1 for the true alternative (because under those priors, different DAGs represented by the same pattern will typically all have non-zero posterior probabilities.) A Bayesian point estimate of the effect of a manipulation can be obtained by taking the expected value (under the posterior distribution) of the effects predicted by each of the causal DAGs and its associated parameters. In this way, a Bayesian numerical point estimate of the effect of any manipulation can be obtained for any data set. However, a Bayesian point estimate of the effect of a manipulation will not in general be Bayes

consistent, unless the predicted effect of a manipulation is the same for every DAG represented by the true pattern.

The space of alternatives could also be taken to be causal patterns. In that case the posterior probability for the true alternative will converge to 1. A Bayesian point estimate of the effect of a manipulation can be obtained by taking the expected value (under the posterior distribution) of the effects predicted by each of the causal patterns and its associated parameters (where the possibility exists that a “don’t know” answer is obtained for some manipulations and some data sets.) Such an estimator will be Bayes consistent (in the weakened sense that allows for “don’t know” answers.)

There are a number of computational difficulties associated with calculating posterior probabilities over either the space of causal DAGs or the space of causal patterns. Because there are a huge number of possible DAGs, it is a non-trivial problem to assign priors to each causal DAG, and to the parameters for each causal DAG. Heckerman (1998) discusses techniques by which this can be accomplished.

It is also impossible in practice to calculate the posterior probability for a single causal DAG, let alone all causal DAGs. However, techniques have been developed to quickly calculate the ratio of posterior probabilities of any two DAGs. As an approximation of a Bayesian solution, then, it is possible to search among the space of DAGs (or of the space of patterns), and output the DAGs (or patterns) with the highest posterior probabilities. (A variation of this is performing a search over the space of DAGs, but turning each of the DAGs output into the pattern that represents the DAG as a final step, in order to determine whether the point estimate is Bayes consistent.) A wide variety of searches from the machine learning literature have been proposed as search algorithms for locating the DAGs with the highest posterior probabilities. These include simple hill-climbing (at each stage choosing from among all of the DAGs that can be obtained from the current best candidate DAG by a single modification the one with the highest posterior probability), genetic algorithms, simulated annealing, etc. See Spirtes et al. (2000) for a summary.

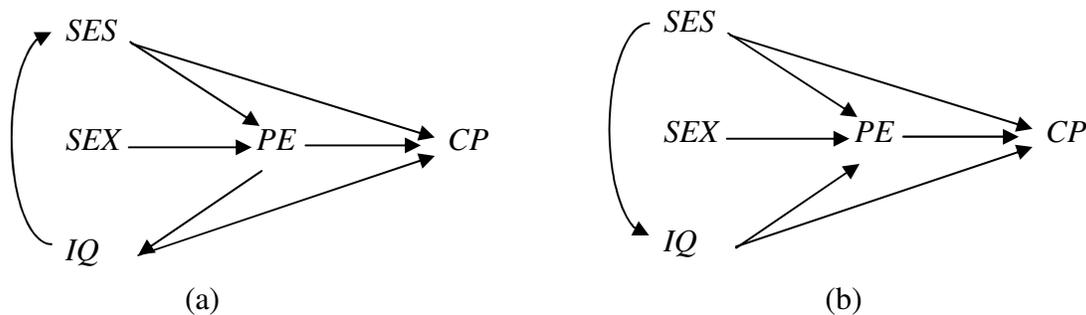


Figure 6: Bayesian Search Output (Assuming No Latent Common Causes)

As an example, consider again the College Plans data. Under the assumption of no latent common causes, that *SEX* and *SES* have no parents, and *CP* has no children, under a variety of different priors, the two DAGs with the highest prior probability are shown in Figure 6(a) and (b). The DAG in Figure 6(b) is the same as the DAG in Figure 4(a). The DAG in Figure 6(a), however, has a posterior probability that is on the order of 10^{10} times more probable than the DAG in Figure 6(b). This is because although the DAG in Figure 6(b) fits the data better, the DAG in Figure 6(a) is much simpler, having only 68 free parameters.

In contrast to the DAG in Figure 6(b),

$$P(CP | PE \parallel P'(PE)) = \sum_{SES} P(CP | PE, SES) \times P(SES)$$

The following are the estimates for $P(CP|PE||P'(PE))$ given the Sewell and Shah data, and assuming the pattern in n Figure 6(a) is correct:

$$\begin{aligned} P(CP=0|PE=0||P'(PE)) &= .080 & P(CP=1|PE=0||P'(PE)) &= .920 \\ P(CP=0|PE=1||P'(PE)) &= .516 & P(CP=1|PE=1||P'(PE)) &= .484 \end{aligned}$$

5.2.7.3. *Score Based Searches*

For computational reasons, the full Bayesian solution of calculating the posterior probability of each DAG, or the posterior probability of the effect of a manipulation cannot be carried out. The approximate Bayesian solution, in effect uses the posterior probability as a way of assigning scores to DAGs, which can then be incorporated into a procedure that searches for the DAGs (or patterns) with the highest score. There are a variety of other scores that have the property that in the large sample limit with probability 1 the true DAG will have a score that is not exceeded by any other DAG. These include for example the Bayesian Information Criterion⁹, which assigns a score that rewards a model for having a high maximum likelihood estimate of the 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.) The Bayes Information Criterion is also a good approximation to the posterior probability in the large sample limit. See Heckerman (1998), and Bollen and Long (1997).

5.2.7.4. *Constraint Based Search*

There are several kinds of searches that in the large sample limit return the correct pattern. The PC algorithm takes as input a covariance matrix or discrete data counts, distributional assumptions, optional background knowledge (e.g. time order), and a significance level, and outputs a pattern. The significance level cannot be interpreted as the probability of type I error for the pattern output, but merely as a parameter of the search. Typically, the significance level is set quite high for small samples sizes (e.g. .15 or .2 for sample size 100), and quite low for large sample sizes (e.g. .01 or .001 for sample size 10000). The search proceeds by performing a sequence of conditional independence tests. In the large sample limit, the search is guaranteed to be correct under the Causal Markov and Causal Faithfulness Principles (if the significance level of the tests performed approaches zero as the sample size approaches infinity). The length of time that the algorithm takes to run depends upon how many parents each variable has. In the worst case (where some variable has all the other variables as parents) the time it takes to perform the search grows exponentially as the number of variables. However, in some cases, it can perform searches on up to 100 measured variables. How large a set of SEMs is represented by the output depends upon what the true SEM is (if such exists).

⁹ The formal definition is given in the Appendix.

One disadvantage of a constraint based algorithm is that it outputs only a single pattern, and gives no indication of whether there are other patterns that explain the data almost as well as the output pattern, but give very different predictions about the effects of manipulations. A partial answer to this problem is to run the algorithm with different significance levels, or to perform a bootstrap test of the output. The output of the PC algorithm on the College Plans data (on significance levels ranging from .001 to .05) is the pattern in Figure 5. This pattern represents the second most probable DAG found in Heckerman (1998) and given the restrictions assumed by Heckerman is the only DAG represented by the pattern. Also, the estimate of $P(CP|PE||P'(PE))$ given by the pattern output by the PC algorithm, and the estimate of $P(CP|PE||P'(PE))$ given by the pattern that represents the best DAG found by Heckerman are fairly close. A bootstrap test of the PC algorithm (with significance level .001) produced the same model as in Figure 5 on 8 out of 10 samples. On the other two samples the edge between PE and CP was not oriented.

Although the set of causal models represented by the pattern in Figure 5 were the best models without latent variables that the PC algorithm could find, the set of conditional independence relations judged to hold in the population are not faithful to any causal model without latent variables. We will discuss relaxing the “no latent variable assumption” imposed by the PC algorithm in section 6.2.7.3.

6. Unmeasured Common Causes

For the parametric families of distributions that we have considered, it is not necessary to introduce latent variables into a model in order to be able to construct uniformly consistent estimators of conditional probabilities. Introducing a latent variable into a model may aid in the construction of consistent estimators that have smaller mean squared error on small samples. This is particularly true of discrete variable models, in which models such as one that has a DAG with one latent variable that is a parent of every measured variable has often proved useful in making predictions.

However, when a model is to be used to predict the effects of manipulations, then the introduction of latent variables into a graph is not merely useful for the sake of constructing low variance estimators, but also for constructing consistent estimators. Unfortunately, as described in this section, latent variables causal models, as opposed to causal models in which the measured variables are causally sufficient, face a number of extra problems that complicate estimation of the effects of manipulations.

6.1. *Known Causal Graph*

In many cases, the parameters of a DAG model with latent variables can still be consistently estimated despite the presence of latent variables. There are a number of algorithms for such estimations, including instrumental variables estimators, and iterative algorithms which attempt to maximize the likelihood. If the model parameters can be estimated, then since the effects of manipulations are functions of the model parameters, the effects of manipulations can also be consistently estimated. ??? However, consistently estimating the model parameters of a latent variable model presents a number of significant difficulties.

1. It is not always the case that the model parameters are functions of the measured variables. This is true of most factor analytic models, for example. In that case, the model parameters are said to be “underidentified”. For parametric families of distributions, whether or not a causal parameter is underidentified is essentially an algebraic problem with a known solution. Unfortunately, algorithms for determining whether a causal parameter is underidentified are too computationally expensive to be run on more than a few variables. There are a number of computationally feasible known necessary conditions for underidentification, and a number of computationally feasible known sufficient conditions for underidentification. See Bollen (1989), Geiger and Meek (1999), and Becker et al. (1994).
2. Even when the model parameters are identifiable, the family of marginal distributions (over the observed variables) associated with a DAG with latent variables lacks many desirable statistical properties that the family of distributions associated with a DAG without measured variables. For example, for SEMs with Normally distributed variables, there are in general no known proofs of the asymptotic existence of maximum likelihood estimators of the model parameters. See Geiger et al. (1999).
3. The estimation of model parameters is often done by iterative algorithms, which are computationally expensive, suffer from convergence problems, and can get stuck in local maxima. See Bollen (1989).

There are also cases where the model parameters are not identifiable, but the effects of some manipulations are identifiable. See Pearl (2000), and Pearl and Robins (1995). A simple example of this is given by application of “the backdoor criterion” (Pearl, 2000) to the model in Figure 7, where X , Y , and Z are binary and measured, and L is ternary and unmeasured. In that case, the model parameters are unidentifiable. However, it can be shown that

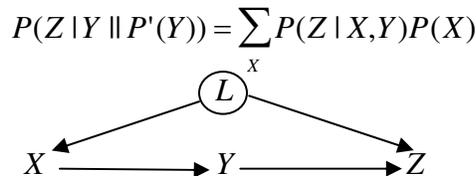


Figure 7: The Backdoor Criterion

The question of whether the effect of a given manipulation is identifiable is for parametric families of distributions an algebraic question whose solutions are known. However, the general algorithms for calculating the solutions are too computationally expensive to be applied to models with more than a few variables. Special cases, which are computationally feasible, are given in Pearl(2001).

6.2. *Unknown Causal Graph*

We will revisit the question of under what conditions and assumptions there are consistent estimators of the effects of manipulations when the measured set of variables may not be causally sufficient.

6.2.1. Distribution and Conditional Independence Equivalence

It is possible that two directed graphs entail exactly the same set of conditional independence relations over a set of measured variables, but are not distributionally equivalent over \mathbf{O} , as long as at least one of them contains a latent variable, a correlated error, or a cycle. For example, consider the DAG in Figure 2 and the DAG in Figure 8. Both entail no conditional independence relations among just the measured variables (*Mechanics*, *Vector*, *Algebra*, *Analysis*, and *Statistics*).

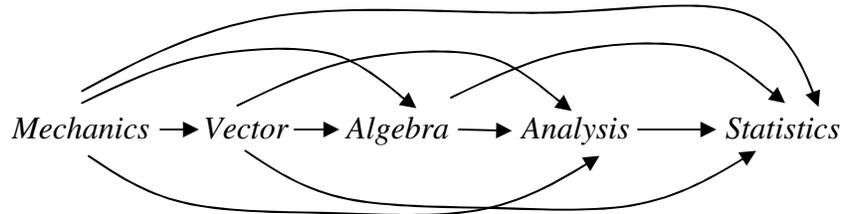


Figure 8: A Complete Graph

However, the SEM in Figure 2 does impose the following non-conditional independence constraints on the correlation matrix among the measured variables:

$$\begin{aligned} \rho(\text{Mechanics}, \text{Algebra}) \times \rho(\text{Vector}, \text{Analysis}) &= \rho(\text{Mechanics}, \text{Analysis}) \times \rho(\text{Vector}, \text{Algebra}) \\ \rho(\text{Mechanics}, \text{Algebra}) \times \rho(\text{Vector}, \text{Statistics}) &= \rho(\text{Mechanics}, \text{Statistics}) \times \rho(\text{Vector}, \text{Algebra}) \\ \rho(\text{Mechanics}, \text{Analysis}) \times \rho(\text{Vector}, \text{Statistics}) &= \rho(\text{Mechanics}, \text{Statistics}) \times \rho(\text{Vector}, \text{Analysis}) \\ \rho(\text{Mechanics}, \text{Analysis}) \times \rho(\text{Algebra}, \text{Statistics}) &= \rho(\text{Mechanics}, \text{Statistics}) \times \rho(\text{Vector}, \text{Analysis}) \\ \rho(\text{Vector}, \text{Statistics}) \times \rho(\text{Algebra}, \text{Analysis}) &= \rho(\text{Vector}, \text{Analysis}) \times \rho(\text{Vector}, \text{Analysis}) \end{aligned}$$

These “vanishing tetrad constraints” were described in Spearman (1904). It is easy to calculate which vanishing tetrad constraints are entailed by a given DAG. Hence these constraints can be useful in searching for causal models with latent variables (Glymour, et al, 1987, Spirtes et al., 2000) as described further in section 6.2.7.4. However, the existence of many kinds of non-conditional independence constraints entailed by a DAG (possibly together with a distribution family) makes the determination of distribution equivalence over \mathbf{O} computationally infeasible.

The DAG in Figure 8 does not impose these constraints on the correlation matrix; it is compatible with any correlation matrix. Thus the two SEMs are conditional independence equivalent over \mathbf{O} (i.e. they entail exactly the same set of conditional independence relations among variables in \mathbf{O}), but they are not distributionally equivalent over \mathbf{O} .

Although it is theoretically possible to determine when two SEMs or two Bayesian networks with latent variables are distributionally equivalent over \mathbf{O} , or to find features common to a distributional equivalence class over \mathbf{O} , in practice it is not computationally feasible (Geiger and Meek 1999) for models with more than a few variables. In addition, if an unlimited number of latent variables are allowed, the number of DAGs that are distributionally equivalent over \mathbf{O} may be infinite.

This implies that the strategy that was used to estimate the effects of manipulations when there were no latent variables cannot be carried forward unchanged to the case where there may be latent variables. There is, however, a modification of that strategy, which is not as

informative as the computationally infeasible strategy of using distribution equivalence classes, but is nevertheless correct.

If \mathbf{O} represents the set of measured variables in path diagrams G_1 and G_2 , then G_1 and G_2 are **conditional independence equivalent over \mathbf{O}** if and only if they entail the same set of conditional independence relations among the variables in \mathbf{O} . It is often far from obvious what constitutes a complete set of DAGs distribution equivalent to a given graph (Spirtes et al. 1998). We will call such a complete set a **conditional independence equivalence class over \mathbf{O}** . (If it is a complete set of graphs without correlated errors or directed cycles, i.e. directed acyclic graphs, that are distribution equivalent we will call it a **simple conditional independence equivalence class over \mathbf{O}** .) A conditional independence equivalence class over \mathbf{O} is always a superset of a distribution equivalence class over \mathbf{O} . For example, suppose $\mathbf{O} = \{SES, SEX, PE, CP, IQ\}$. There are only two DAGs without latent variables in the conditional equivalence class over \mathbf{O} shown in Figure 4. However, there are other DAGs that have latent variables that are in the same conditional independence equivalence class over \mathbf{O} . Two such DAGs are shown in Figure 9.

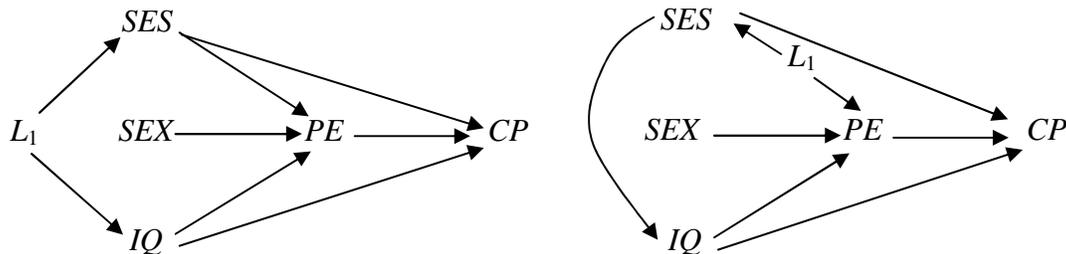


Figure 9: Some Latent Variable Graphical Models in the Simple Conditional Independence Equivalence Class Over \mathbf{O}

A conditional independence equivalence class over \mathbf{O} contains an infinite number of DAGs, because there is no limit to the number of latent variables that may appear in the DAG.

Instead of calculating the effects of manipulations for which every member of the distribution equivalence class over \mathbf{O} agree, we can calculate the effects only of those manipulations for which every member of the conditional independence equivalence class over \mathbf{O} agree. This will typically be fewer manipulations than could be predicted given the distributional equivalence class over \mathbf{O} (because a larger set of graphs have to make the same prediction), but the predictions made will still be correct.

6.2.2. Features Common to a Simple Conditional Independence Equivalence Class

Even though the set S of DAGs in a simple conditional independence equivalence class over \mathbf{O} is infinite, it is still possible to extract the features that the members of S have in common. For example, every member of the simple conditional independence class over \mathbf{O} that contains the DAG in Figure 1 has a directed path from PE to CP and no latent common cause of PE and CP . This is informative because even though the data does not help choose between members of the equivalence class, insofar as the data is evidence for the disjunction of the members in the equivalence class, it is evidence that PE is a cause of CP .

In section 6.2.4 we discuss how to extract features common to a simple conditional independence class over \mathbf{O} .

6.2.3. Conditional Independence Equivalence for DAGs With Latent Variables

Testing whether two DAGs are conditional independence equivalent over \mathbf{O} requires a much more complicated algorithm than does testing whether two DAGs are conditional independence equivalent (or distribution equivalent for multivariate Normal or discrete variables). If V is the maximum of the number of variables in G_1 or G_2 , and M is the number of variables in \mathbf{O} , Spirtes and Richardson (1996) presents an $O(M^3 \times V^2)$ algorithm for checking whether two acyclic path diagrams G_1 and G_2 (which may contain latent variables and correlated errors) are d-separation equivalent over \mathbf{O} .

6.2.4. Extracting Features Common to a Simple conditional Independence Equivalence Class

There is a graph called a partial ancestral graph, analogous to a pattern, which represents the features common to a simple conditional independence over \mathbf{O} equivalence class. The partial ancestral graph over \mathbf{O} for the DAG in Figure 1 is shown in Figure 10. Two variables A and B are adjacent when they are dependent conditional on every subset of the variables in $\mathbf{O} \setminus \{A, B\}$. The “-” endpoint of the $PE \rightarrow CP$ edge means that PE is an ancestor of CP in every DAG in the simple equivalence class over \mathbf{O} . The “>” endpoint of the $PE \rightarrow CP$ edges means that CP is not an ancestor of PE in any member of the equivalence class. The “o” endpoint of the $SES \rightarrow CP$ edge makes no claim about whether SES is an ancestor of CP or not. The line connecting the $SES \rightarrow CP$ edge and the $SES \rightarrow PE$ edge means that in every DAG in the equivalence class, SES is an ancestor of either PE or CP .

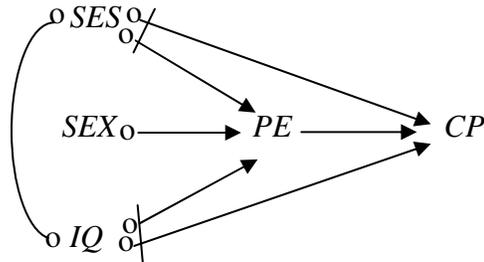


Figure 10: Partial Ancestral Graph

The FCI algorithm, discussed in section 6.2.7.3, can be used to construct a PAG from a given DAG.

6.2.5. Calculating the Effects of Manipulations from a PAG

The rules that specify which effects of manipulations can be calculated from a PAG and how to calculate them, and which effects of manipulations cannot be calculated from a PAG are described in Spirtes, Glymour, and Scheines (2001). Here we give some examples without proof.

Suppose that it is known that the PAG in Figure 10 is the correct PAG. In this case the simple conditional independence over \mathbf{O} equivalence class S is a superset of the

corresponding simple conditional equivalence class T (i.e. the DAGs which contain members of \mathbf{O} but not any latent variables.) Hence, there can be no more predictions of effects of manipulation common to every member of S that there are predictions of effects of manipulations common to every member of T (and in this case there are fewer.) As in the case where it is assumed that the pattern is given, if the PAG is given there is no consistent estimator of $P(IQ \parallel P'(SES))$. Different members of the simple conditional independence over \mathbf{O} equivalence class (e.g. the DAGs in Figure 4) predict different effects.

In contrast to the case where it is assumed that the pattern is given, if the PAG is given, there are no consistent estimators of $P(PE \mid SES, SEX \parallel P'(PE))$.

As in the case where it is assumed that the pattern is given, if the PAG is given there is a uniformly consistent estimator of $P(CP \mid PE \parallel P'(PE))$ because in each DAG represented by the PAG

$$P(CP \mid PE \parallel P'(PE)) = \sum_{IQ, SES} P(CP \mid PE, SES, IQ) \times P(IQ \mid SES) \times P(SES)$$

6.2.6. Bayes, Pointwise, and Uniformly Consistent Estimators

Suppose that the only given data are samples from a multivariate Normal probability distribution, or a distribution over all discrete variables, and that the set of variables is not known to be causally sufficient. Under what assumptions and conditions are there pointwise or uniformly consistent estimators of manipulations that are functions of the sample, when the causal DAG is not given? The answers are provided in Table 2. Note that the only line that has changed from Table 1 is the last line, where a time order is assumed. Unlike the case of known causal sufficiency, a known time order does not entail the existence of consistent estimators.

Time order	Causal Faithfulness Prior	Causal Faithfulness	Strong Causal Faithfulness	Existence of Bayes Consistent	Existence of Pointwise Consistent	Existence of Uniformly Consistent
NO	NO	NO	NO	NO	NO	NO
NO	YES	NO	NO	YES	NO	NO
NO	YES	YES	NO	YES	YES	NO
NO	YES	YES	YES	YES	YES	YES
YES	NO	NO	NO	NO	NO	NO

Table 2

However, in each case, when the possibility of latent common causes is allowed, there are more cases in which the consistent estimators returns “can’t tell” than if it is assumed that there are no latent common causes. See Spirtes et al. (2001), Robins (forthcoming), and Zhang (2002).

6.2.7. Model Selection and Construction

Even if the causal process that generated the College Plans data had an unmeasured variable, there are uniformly consistently estimators of $P(CP \mid SEX, IQ, SES, PE)$ from models

without latent variables, e.g. using the relative frequency of CP conditional on SEX , IQ , SES , and PE . A model with latent variables may provide a more efficient estimate of $P(CP|SEX,IQ,SES,PE)$, but it is not needed in order to achieve uniform consistency.

On the other hand, if the model is to be used to understand the processes by which CP are determined, or to predict the effect on CP of a manipulation of PE (i.e. $P(CP||P'(PE))$), then a very different methodology is called for, and the possibility of latent variables must be accounted for in order to construct a uniformly consistent estimator $P(CP||P'(PE))$. Again, consistent estimators can be constructed in a two-step process. First, perform a search that in the large sample limit returns the correct conditional independence equivalence class over \mathbf{O} . Second, if the effect of a manipulation is to be calculated, calculate it from the conditional independence equivalence class over \mathbf{O} and the sample distribution (as described in section 6.2.5.) Here we will describe searches that return the correct conditional independence equivalence class over \mathbf{O} in the large sample limit.

6.2.7.1. Bayesian Causal Inference

As in the case of causal DAGs without latent variables, in practice Bayesian inference for causal DAGs with latent variables is computationally infeasible. Hence, instead of calculating the posterior probabilities of each causal DAGs, the ratio of posterior probabilities are calculated, and a small set of causal DAGs with the highest posterior probabilities are output.

However, introducing latent variables into a causal model also introduces two new major problems for a Bayesian search strategy. The first problem is that the number of possible causal DAGs is infinite, so calculating the posterior probability for each of them is out of the question. The second problem is that even with very simple priors, calculating the ratio of posterior probabilities for two causal DAGs with latent variables is computationally much more expensive than calculating the ratio of posterior probabilities for two causal DAGs without latent variables. Various approximations can be used to simplify the calculations (Heckerman 1998).

Heckerman (1998) describes a Bayesian search over a number of different latent variable models. The best model that they found for the College Plans data is shown in Figure 11, where H is a latent variable. It is more than 10^{10} times more likely than the best model with no latent variables, and more than 10^9 times more likely than the next best model with latent variables. The DAG in Figure 11 is represented by the same PAG as the PAG that represents the best DAG model without latent variables (the DAG in Figure 6(a)). Also, $P(CP|PE||P'(PE))$ is the same for both DAG models.

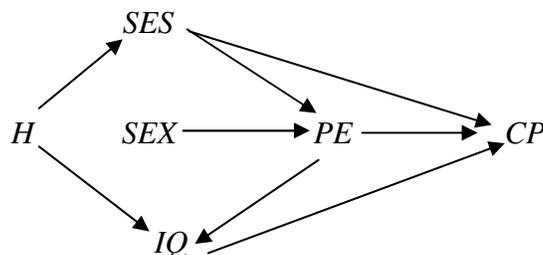


Figure 11: Output of Bayesian Model Search (with Latent Variables)

6.2.7.2. Score Based Searches

Score based searches in latent variable models face similar obstacles to Bayesian searches. The search space is infinite. Calculating the dimension of a latent variable model is computationally quite difficult; in general the concept of the complexity for a latent variable model is not well defined, as the dimensionality of the model can be different in different parts of the parameter space (Geiger et al. 1999.) Even for multivariate Normal or discrete latent variable models, the existence of maximum likelihood estimates in the large sample limit has not been demonstrated. It has not been shown that the Bayes Information Criterion, which is a good approximation to the posterior in the large sample limit for DAG models without latent variables has the same property for latent variable DAG models.

Friedman (1997) describes a heuristic search that interleaves estimation steps with steps that modify the structure of the model to improve the fit.

6.2.7.3. Constraint Based Search

The FCI algorithm is a more complicated version of the PC algorithm that does not make the assumption that there are no latent causes of pairs of observed variables, and hence returns PAGs instead of patterns. It returns the true PAG in the large sample limit, assuming the Causal Markov and Causal Faithfulness Principles. The FCI algorithm takes as input a sample, distributional assumptions, optional background knowledge (e.g. time order), and a significance level, and outputs a PAG. One advantage of searching the space of PAGs, instead of the space of DAGs with latent variables is that the former is finite, while the latter is not. In addition, because the algorithm needs only tests of conditional independence, it avoids the computational problems involved in calculating posterior probabilities or scores for latent variable models.

The output of the FCI algorithm on the College Plans data (run at significance levels varying from .001 to .05) is the PAG shown in Figure 12. Note that it is not the same as the PAG that represents the causal models output by the PC algorithm (as represented by the pattern in Figure 5). The reason for this is that the PAG that most closely represents the conditional independence relations judged to hold in the population, represents only causal models with latent variables (as evidenced by the double-headed arrows in the PAG.)

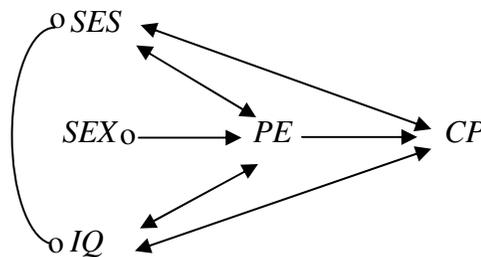


Figure 12: Output of the FCI Algorithm

If the PAG in Figure 12 is correct, then $P(CP|PE||P'(PE)) = P(CP|PE)$. From the Sewell and Shah data, the estimate of $P(CP|PE||P'(PE))$ is:

$$\begin{array}{ll}
 P(CP=0|PE=0||P'(PE)) = .063 & P(CP=1|PE=0||P'(PE)) = .937 \\
 P(CP=0|PE=1||P'(PE)) = .572 & P(CP=1|PE=1||P'(PE)) = .428
 \end{array}$$

Again, these estimates are close to the estimates given by the output of the PC algorithm, and the Bayesian search algorithm. A bootstrap test of the output run at significance level 0.001 yielded the same PAG as in Figure 12 on 8 out of 10 samples. In the other two samples, the edge between *PE* and *CP* was unoriented.

6.2.7.4. Multiple Indicator Models

The model shown in Figure 2 is an example of a multiple indicator model. Multiple indicator models can be divided into two parts. The causal relationships between the latent variables are called the structural model. The rest of the model is called the measurement model. The structural model is $Vector\ Algebra\ Skill \leftarrow Algebra\ Skill \rightarrow Real\ Analysis\ Skill$, and the measurement model consists of the graph with all of the other edges. Typically, the structural model is the part of the model that is of interest.

The FCI algorithm can be applied to the measured variables in a multiple indicator model, but the results, while correct in the large sample limit, are not informative. Assuming the Causal Faithfulness Principle, in the large sample limit, for each pair of measured variables X and Y , the PAG that is output contains an undirected edge $X \text{ o---o } Y$. It is not possible to predict the effects of any manipulation from such a PAG, and the algorithm that constructs it is exponential in the number of variables. (The sample size for the Mathematics Marks data is quite small (88), and the actual PAG that is output is from a pattern of conditional independence relations that would be entailed by the DAG in Figure 2 if *Algebra* was a very good measure of *Algebra Skill*. It still fails to be informative about the relationships of interest, namely what the causal relationships between the latent variables are.)

However, if the measurement model is known from background knowledge, then the measurement model can be used to perform tests of conditional independence among the latent variables. For example, to test whether $\rho(Vector\ Algebra\ Skill, Real\ Analysis\ Skill \mid Algebra\ Skill) = 0$, a chi-squared test comparing the model in Figure 2 with the model which differs only by the addition of an edge between *Vector Algebra Skill* and *Real Analysis Skill* can be performed. If the difference between the two models is not significant then the partial correlation is zero. Thus the FCI or PC algorithm can be applied directly to discover the structural model. At a significance level of .2, the FCI algorithm produces the PAG $Vector\ Algebra\ Skill \text{ o---o } Algebra\ Skill \text{ o---o } Real\ Analysis\ Skill$, which is the PAG that represents the structure model part of the model shown in Figure 2.

A variety of degrees of background knowledge about the measurement model are possible. For example, background knowledge might indicate for each measured variable which latent variable caused it, but not whether there might also be other causes of the measured variable as well. For example, it might be known that *Vector Algebra Skill* causes *Mechanics* and *Vector*, and *Real Analysis Skill* causes *Analysis* and *Statistics*. But it might not be known whether the measurement model is “pure”¹⁰, i.e. whether there are no other latent common causes of *Analysis* and *Statistics*, or of *Mechanics* and *Vector*, and no causal relations among the measured variables. However, the constraint that

¹⁰ More formally, a measurement model is “pure” if each measured variable is d-separated from all of the other measured variables conditional on the latent variable that it measures.

$$\rho(\text{Mechanics,Analysis}) \times \rho(\text{Vector,Statistics}) = \rho(\text{Mechanics,Statistics}) \times \rho(\text{Vector,Analysis})$$

is entailed by the graph only if the measurement model is pure. So it is possible to use a test of the vanishing tetrad constraint to determine whether a measurement model is pure, or if it requires additional edges. If there are many indicators of each latent, tests of the vanishing tetrad constraints can also be used to select a subset of the indicators that form a pure measurement model. See Anderson and Gerbing (1987 and 1988) on pure measurement models, and Spirtes et al. (2000) on using vanishing tetrad constraints for selecting pure measurement models.

If the number of latents is not known, or it is not known which measured variables are indicators of which latents, it might be hoped that factor analysis could be used to create a correct measurement model. However, Glymour (1997) describes some simulation experiments in which factor analysis does not do well at constructing measurement models.

6.2.7.5. *Distribution and Conditional Independence Equivalence for Path Diagrams with Correlated Errors or Directed Cycles*

The representation of feedback using cyclic graphs, and the theory of inference of cyclic graphs from data is not as well developed as for the case of DAGs, except in special cases. There are general algorithms for testing distribution equivalence for multivariate Normal graphical models with correlated errors or directed cycles, but they are generally computationally infeasible for more than a few variables (Geiger and Meek, 1999).??? For multivariate Normal variables, Spirtes (1994) and Koster (1995) proved that all of the conditional independence relations entailed by a graph with correlated errors and cycles are captured by the (natural extension of) the d-separation relation to cyclic graphs, and Pearl and Dechter (1996) proves an analogous result for discrete variables. However, Spirtes (1994) proved that given non-linear relationships among continuous variables, it is possible for \mathbf{X} to be d-separated from \mathbf{Y} conditional on \mathbf{Z} , but for \mathbf{X} and \mathbf{Y} to be dependent conditional on \mathbf{Z} . There are computationally feasible algorithms for testing conditional independence equivalence for multivariate Normal graphical models with correlated errors or directed cycles, but no latent variables, and extensions of the PC algorithm to multivariate Normal or discrete variable graphs with cycles (Richardson, 1996), but there is no known algorithm for inferring graphs with both cycles and latent variables. Lauritzen and Richardson (2002) discusses the representation of one kind of feedback using not cyclic graphs, but an extension of DAGs called chain graphs.

7. Some Common Social Science Practices

In this section we will examine the soundness of several practices in the social sciences commonly used to draw causal inferences.

7.1. *The Formation of Scales*

It is a common practice when attempting to discover the causal relations between latent variables to take all of the indicators of a given latent variable and average them together to form a “scale”. The scale variable is then substituted for the latent variable in the analysis.

The following simulated example shows why this practice does not yield reliable information about the causal relationships between latent variables.

For the model in Figure 13, hereafter the “true model,” 2,000 data points were generated. All exogenous variables error terms are independent standard normal variates.

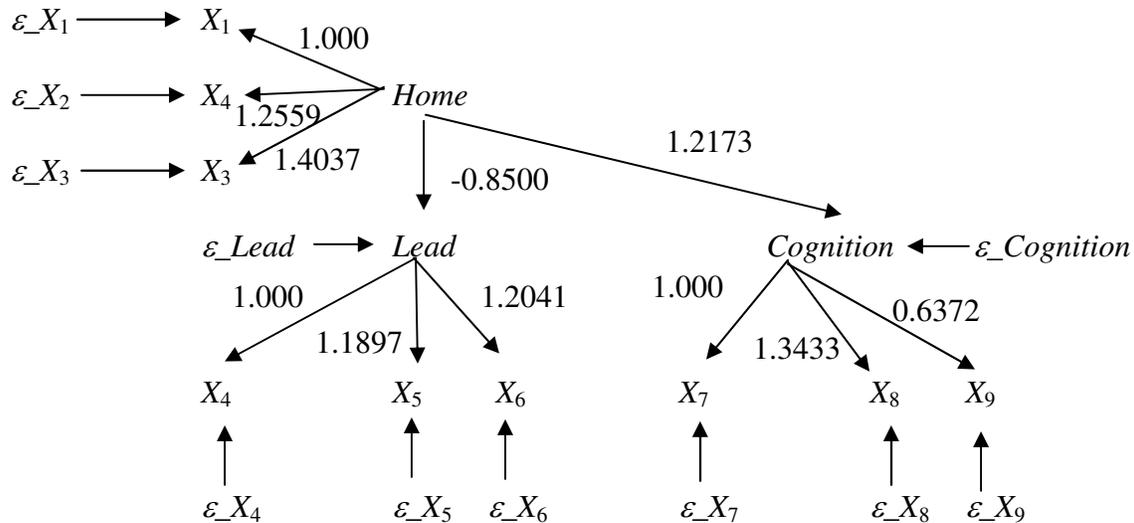


Figure 13: Simulated Lead Study

Suppose the question under investigation is the effect of *Lead* on *Cognition*, controlling for the *Home* environment. Given the true model, the correct answer is 0 - that is, *Lead* has *no direct effect* on *Cognition* according to this model. Consider first the ideal case in which we suppose that we can directly and perfectly measure *Home*, *Lead* and *Cognition*. To test the effect of *Lead* on *Cognition*, we might regress *Cognition* on *Lead* and *Home*. Finding that the linear coefficient on *Lead* is $-.00575$, which is insignificant ($t = -.26$, $p = .797$), we correctly conclude that *Lead's* effect is insignificant.

Second, consider the case in which *Lead* and *Cognition* were directly measured, but *Home* was measured with a scale that averaged X_1 , X_2 , and X_3 , the indicators of *Home* in the true model: $Homescale = (X_1 + X_2 + X_3)/3$. Regressing *Cognition* on *Lead* and controlling for *Home* with the *Homescale* variable and finding that the coefficient on *Lead* is now $-.178$, which is significant at $p = .000$, we incorrectly conclude that *Lead's* effect on *Cognition* is deleterious.

Third, consider the case in which *Lead*, *Cognition*, and *Home* were all measured with scales: $Homescale = (X_1 + X_2 + X_3)/3$, $Leadscale = (X_4 + X_5 + X_6)/3$, $Cogscale = (X_7 + X_8 + X_9)/3$. Regressing *Cogscale* on *Homescale* and *Leadscale* gives a coefficient on *Leadscale* of $-.109$, which is still highly significant at $p = .000$, so we would again incorrectly conclude that *Lead's* effect is deleterious.

Next consider a strategy in which we build a scale for *Home* as we did above, i.e., $Homescale = (X_1 + X_2 + X_3)/3$, and use it in place of the latent *Home* and its indicators X_1 , X_2 , and X_3 . In one important respect the result is worse. In this case the estimate for the effect of *Lead* on *Cognition*, controlling for the home environment (*Homescale*) is $-.137$, which is highly significant at $t = -5.271$, and thus substantively in error. The model as a

whole fits quite well ($\chi^2 = 14.57$, $df = 12$, $p = .265$), and all the distributional assumptions are satisfied, so nothing in the statistical treatment of this case would indicate that we have misspecified the model, even though the estimate of the influence of *Lead* is quite incorrect.

7.2. Regression

It is common knowledge among practicing social scientists that in order for the coefficient of X in the regression of Y upon X to be interpretable as the effect of X on Y there should be no "confounding" variable Z that is a cause of both X and Y . See Figure 14.

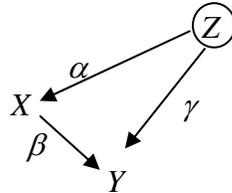


Figure 14: The Problem of Confounding

The coefficient from the regression of Y on X alone will be a consistent estimator only if either α or γ is equal to zero. Further, observe that the bias term $\alpha\gamma V(Z)/V(X)$ (where $V(Z)$ is the variance of Z) may be either positive or negative, and of arbitrary magnitude. However, the coefficient of X in the regression of Y on X and Z is a consistent estimator of β since $\text{Cov}(X, Y|Z)/V(X|Z) = \beta$.

The danger presented by failing to include confounding variables is well understood by social scientists. Indeed, it is often used as the justification for considering a long "laundry list" of "potential confounders" for inclusion in a given regression equation. What is perhaps less well understood is that including a variable that is not a confounder can also lead to biased estimates of the structural coefficient. We now consider a simple example demonstrating this. In this example, Z may temporally precede both X and Y .

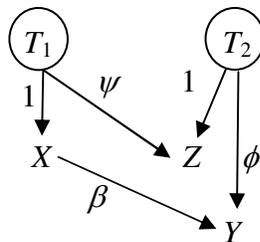


Figure 15: Estimates Biased by Including More Variables

In the path diagram depicted in

Figure 15 there are two unmeasured confounders T_1 and T_2 , which are uncorrelated with one another. It can be shown that the coefficient of X in the regression of Y on X and Z is not a consistent estimate of β , (unless $\rho(X, Z) = 0$ or $\rho(Y, Z) = 0$), and may even have a completely different sign. In the case where $\beta = 0$, the coefficient of X in the regression of Y on X will be zero in the population, but will become non-zero once Z is included.

SEM folklore often appears to suggest that it is better to include rather than exclude a variable from a regression (barring statistical problems such as small sample size) If the goal of a model is to predict the value of a hidden variable, rather than the result of a

manipulation, this is sound advice (ignoring for the moment such statistical problems as small sample size, or collinearity.) However, if the goal of a model is to describe causal relations, or to predict the effects of a manipulation, this is not a theoretically sound practice. The notion that adding more variables is always a sound practice perhaps given support by reference to “controlling for Z ”, the implication being that controlling for Z eliminates a source of bias. The conclusion to be drawn from these examples is that there is no sense in which one is “playing safe” by including rather than excluding “potential confounders”; if they turn out not to be potential confounders then this could change a consistent estimate into an inconsistent estimate. Of course, this does not mean that on average, that one is not better off regressing on more variables than fewer: whether or not this is the case depends upon the distribution of the parameters in the domain. Greenland ???

The situation is also made somewhat worse by the use of misleading definitions of 'confounder': sometimes a confounder is said to be a variable that is strongly correlated with both X and Y , or even a variable whose inclusion changes the coefficient of X in the regression. Since, for sufficiently large $\rho(X,Z)$ or $\rho(Y,Z)$, Z in

Figure 15 would qualify as a confounder under either of these definitions, it follows that under either definition including confounding variables in a regression may make a hitherto consistent estimator inconsistent.

If Y is regressed on a set of variables \mathbf{W} , including X , in which SEMs will the partial regression coefficient of X be a consistent estimate of the structural coefficient β associated with the $X \rightarrow Y$ edge? The coefficient of X is a consistent estimator of β if \mathbf{W} does not contain any descendant of Y in G , and X is d-separated from Y given \mathbf{W} in the DAG formed by deleting the $X \rightarrow Y$ edge from G .¹¹ If this condition does not hold, then for almost all instantiations of the parameters in the SEM, the coefficient of X will fail to be a consistent estimator of β . It follows directly from this that (almost surely) β cannot be estimated consistently via any regression equation if either there is an edge $X \leftrightarrow Y$ (i.e. ε_X and ε_Y are correlated) or if X is a descendant of Y (so that the path diagram is cyclic).

8. Appendix

An **undirected path** in a DAG G from X_1 to X_n is a sequence of vertices $\langle X_1, \dots, X_n \rangle$ in which for each pair of vertices X_i and X_{i+1} adjacent in the sequence, there is either an edge $X_i \rightarrow X_{i+1}$ or an edge $X_i \leftarrow X_{i+1}$ in G . A vertex X_i is a **collider** on an undirected path U in G if and only if there are edges $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ on U in G . B is a **descendant** of A in G if and only if either there is a directed path from A to B in G , or $A = B$. If \mathbf{X} , \mathbf{Y} and \mathbf{Z} are disjoint subsets of variables in G , \mathbf{X} and \mathbf{Y} are **d-connected** conditional on \mathbf{Z} if and only if U is an undirected path from some $X \in \mathbf{X}$ to some $Y \in \mathbf{Y}$ such that every collider on U has a descendant in \mathbf{Z} , and every non-collider on U is not in \mathbf{Z} . If \mathbf{X} , \mathbf{Y} and \mathbf{Z} are three disjoint

¹¹Note this criterion is similar to Pearl's back door criterion (Pearl, 1993), except that the back-door criterion was proposed as a means of estimating the *total* effect of X on Y .

sets of vertices, \mathbf{X} and \mathbf{Y} are **d-separated** conditional on \mathbf{Z} in G if and only if there is no undirected path U in G that d-connects \mathbf{X} and \mathbf{Y} conditional on \mathbf{Z} .

Let $\Omega(G)$ be the set of probability distributions that are faithful to G , and $T(P,G)$ be some causal parameter (that is a function of the distribution P and the DAG G .) Let $\Omega\mathcal{G} = \{(P,G):G \in \mathcal{G}, P \in \Omega(G)\}$, $\hat{\theta}$ is **pointwise consistent** if for all $(P,G) \in \Omega\mathcal{G}$, for every $\varepsilon > 0$, $\lim_{n \rightarrow \infty} P^n(d[\hat{\theta}_n(O^n), T(P,G)] > \varepsilon) = 0$. $\hat{\theta}$ is **uniformly consistent** if for every $\varepsilon > 0$, $\lim_{n \rightarrow \infty} \sup_{(P,G) \in \Omega\mathcal{G}} P^n(d[\hat{\theta}_n(O^n), T(P,G)] > \varepsilon) = 0$

The Bayesian Information Criterion is defined as $\log P(D | \hat{\theta}_G, G) - \frac{d}{2} \log N$

where D is the sample data, G is a directed acyclic graph, $\hat{\theta}_G$ is the vector of maximum likelihood estimates of the parameters for DAG G , N is the sample size, and d is a measure of the complexity of the model, which in DAGs without latent variables is simply the number of free parameters in the model. Because $P(G|D) \propto P(D|G)P(G)$, if $P(G)$ is the same for each DAG, the BIC score approximation for $P(D|G)$ can be used as a score for $P(G|D)$.

9. Guide to Literature

In this section, we give a guide to the literature with an emphasis on works that provide broad overviews of causal inference. It is not a guide to the history of the subject.

Pearl (2001), Spirtes et al. (2001), and Lauritzen (2001) provide overviews of the directed graphical approach to causal inference described in this article. Glymour and Cooper (1998) is a collection of articles that also covers many issues about causal inference. Robins (1986) describes a non-graphical approach to causal inference based on Rubin's (1977) counterfactual approach.

Applications of causal inference algorithms are described in Glymour and Cooper (1999) and Spirtes et al. (2001). Biological applications are described in Shipley (2000). Some applications to econometrics are described in Swanson and Granger (1997), and the concept of causality in economics is described in Hoover (2001). The mathematical marks example is from Mardia, Kent, and Bibby (1979), and is analyzed in Whittaker (1990) and Spirtes et al. (2001). The college plans example is analyzed in Sewell and Shah (1968), Spirtes et al. (2001), and Heckerman (1998).

Bollen (1989), Blalock (1971), and Goldberger and Duncan (1973) are introductions to the theory of structural equation models. Sullivan and Feldman (1979) is an introduction to multiple indicator models. Lawley and Maxwell (1971) describes factor analysis in detail. Bartholomew and Knott (1999) is an introduction to a variety of different kinds of latent variable models. Pearl (1988), Neapolitan (1990), Cowell (1999) and Jensen (2001) are introduction to Bayesian networks. Whittaker (1990), Lauritzen (1996), and Edwards (2000) describe a number of different kinds of graphical models.

An overview of machine learning techniques is given in Mitchell(1997). Bollen and Long (1993) describes a number of different methods of evaluating models. A collection of articles about learning graphical models is in Jordan (1998). The Bayesian approach to causal inference is described in detail in Heckerman (1998). Buntine (1996) is an overview of different approaches to search over Bayesian networks. There are also many articles on this subject in the Proceeding of the Conference on Uncertainty in Artificial Intelligence (<http://www2.sis.pitt.edu/~dsl/UAI/>), and the Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (<http://research.microsoft.com/conferences/aistats2003/proceedings/index.htm>).

The results on consistency are in Robins, et al. (forthcoming), Spirtes et al. (2001), and Zhang (2002).

Philosophical works about the meaning and nature of causation are Cartwright (1989 and 1999), Eells (1991), Hausman (1998), Shafer (1996), and Sosa and Tooley(1993).

There are a number of statistical packages devoted to estimating and testing structural equation models. These include the commercial packages EQS (<http://www.mvsoft.com/>), LISREL (<http://www.ssicentral.com/lisrel/mainlis.htm>) and CALIS, which is part of SAS (<http://www.sas.com>). EQS and LISREL also contain some search algorithms for modifying a given causal model. The statistical package R (<http://www.r-project.org/>) also contains a “sem” package for estimating and testing structural equation models.

HUGIN (<http://www.hugin.com/>) is a commercial program that aids in the construction and estimation of Bayesian networks, and also contains algorithms for calculating conditional probabilities using Bayesian networks. (A free demo version is also available.) It also contains an implementation of a modification of the PC algorithm. MSBNx (<http://www.research.microsoft.com/adapt/MSBNx/>) is free software for the creation, estimation, evaluation, and use of Bayesian networks. MIM (<http://www.hypergraph.dk/>) and COCO (<http://www.math.auc.dk/~jhb/CoCo/>) model several kinds of graphical models. BUGS (<http://www.mrc-bsu.cam.ac.uk/bugs/>) is a free program that uses Bayesian techniques to estimate and test graphical models. TETRAD IV (<http://www.phil.cmu.edu/projects/tetrad/>) is a free program contains a number of search algorithms, including the PC and FCI algorithms.

10. References

- Anderson, J., & Gerbing, D. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103, 411-423.
- Anderson, J., and Gerbing, D., & Hunter, J. (1987). On the assessment of unidimensional measurement: internal and external consistency and overall consistency criteria. *Journal of Marketing Research*, 24, 432-437.
- Andersson, S., Madigan, D., and Perlman, M. (1995) A Characterization of Markov Equivalence Classes for Acyclic Digraphs, Technical Report 287, Department of Statistics, University of Washington.
- Bartholomew, D., and Knott, M. (1999) *Latent Variable Models and Factor Analysis*, Edward Arnold, 2nd edition, London, England.

- Becker, P., Merckens, A., and Wansbeek, T. (1994). *Identification, equivalent models, and Computer Algebra*. Academic Press, San Diego, CA.
- Bickel, P., and Doksum, K. (2001) *Mathematical Statistics: Basic Ideas and Selected Topics*, Prentice-Hall, N.J.
- Blalock, H., 1961, *Causal Inferences in Nonexperimental Research*, (W. W. Norton and Co., New York).
- Blalock, H. (ed.) (1971) *Causal Models in the Social Sciences*, Aldine, Chicago.
- Becker, P., Merckens, A., and Wansbeek, T. (1994). *Identification, equivalent models, and Computer Algebra*. Academic Press, San Diego, CA.
- Bollen, K., (1989) *Structural Equations with Latent Variables*. (Wiley, New York).
- Bollen, K. and Long, J. (1993) *Testing Structural Equation Models*. Sage, Newbury Park, CA.
- Buntine, W. (1996). A guide to the literature on learning graphical models. *IEEE Transactions on Knowledge and Data Engineering*, **8**, pp. 195-210.
- Cartwright, N. (1989). *Nature's Capacities and their Measurement*. Oxford, New York, Clarendon Press; Oxford University Press.
- Cartwright, N. (1999) *The Dappled World: A Study of the Boundaries of Science*, Cambridge University Press, New York.
- Chickering, D. (1995) A Transformational Characterization of Equivalent Bayesian Network Structures, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Philippe Besnard and Steve Hanks (Eds.), Morgan Kaufmann Publishers, Inc., San Mateo, CA.
- Cowell, R. (ed.), (1999) *Probabilistic Networks and Expert Systems* (Statistics for Engineering and Information Science), Springer-Verlag, New York.
- Edwards, D. (2000) *Introduction to Graphical Modelling*, 2nd ed., Springer-Verlag, New York.
- Eells, E. (1991) *Probabilistic Causality*, Cambridge University Press, New York.
- Friedman, N. (1997) Learning Bayesian Networks in the Presence of Missing Values and Hidden Variables, in *Fourteenth International Conference on Machine Learning*.
- Geiger, D., Heckerman, D., King, H., and Meek, C. 1999. On the geometry of DAG models with hidden variables. *Artificial Intelligence and Statistics 99*. D. Heckerman and J. Whittaker. San Francisco, CA, Morgan Kauffman.
- Geiger, D. and Meek, C. 1999. On solving statistical problems with quantifier elimination, Microsoft Research.
- Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987). *Discovering Causal Structure*. San Diego, CA, Academic Press.
- Glymour, C. and Cooper, G. (eds.) (1999) *Computation, Causation and Discovery*. MIT Press, Cambridge, MA.
- Glymour, C. (1997) Social Statistics and Genuine Inquiry: Reflections on the Bell Curve, in *Intelligence, Genes and Success*, edited by B. Devlin, S. Fienberg, D. Resnick, and K. Roeder, Springer-Verlag, New York, pp. 257-280.
- Goldberger, A., Duncan, O. (eds.), 1973, *Structural Equation Models in the Social Sciences* (Seminar Press, New York).
- Hoover, K. (2001) *Causality in Macroeconomics*, Cambridge University Press, New York.
- Hausman, D. (1998) *Causal Asymmetries*, Cambridge University Press, New York.
- Heckerman, D. (1998). A tutorial on learning with Bayesian networks. *Learning in Graphical Models*, edited by M. Jordan. MIT Press, Cambridge, MA.

- Jensen, F. (2001) *Bayesian Networks and Decision Graphs* (Statistics for Engineering and Information Science), Springer-Verlag, New York.
- Koster, J., (1995) Markov Properties of Non-Recursive Causal Models, *Annals of Statistics*, November 1995.
- Lauritzen, S. (1996). *Graphical Models*, Oxford University Press, Oxford.
- Lauritzen, S., Dawid, A., Larsen, B., Leimer, H., 1990, Independence properties of directed Markov fields, *Networks*, **20**, 491-505.
- Lauritzen, S. (2001) "Causal Inference from Graphical Models", in *Complex Stochastic Systems*, edited by O. Barnsdorff-Nielsen, D. Cox, and C. Kluppenberg, Chapman and Hall, London, pp. 63-107.
- Lauritzen, S. and Richardson, T. (2002) Chang graph models and their causal interpretation (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, pp. 321-361.
- Lawley, D., and Maxwell, A. (1971). *Factor Analysis as a Statistical Method*. Butterworth, London.
- Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. New York, Academic Press.
- Meek, C. (1995) Causal inference and causal explanation with background knowledge, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Philippe Besnard and Steve Hanks (Eds.), Morgan Kaufmann Publishers, Inc., San Mateo, CA, pp. 403-410.
- Mitchell, T. (1997) *Machine Learning*. WCB/McGraw-Hill, Cambridge, MA.
- Neapolitan, R. (1990). *Probabilistic Reasoning in Expert Systems*. New York, Wiley.
- Needleman, H., Geiger, S., and Frank, R. (1985). "Lead and IQ Scores: A Reanalysis," *Science*, 227, pp. 701-704.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufman, San Mateo, CA.
- Pearl, J. and Dechter, R. (1996). Identifying independencies in causal graphs with feedback. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, CA, pp. 240-246.
- Pearl, J., and Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA. Morgan Kaufmann Publishers, pp. 444-453.
- Pearl, J. (2001) *Causality: Models, Reasoning and Inference*, Cambridge University Press, New York.
- Richardson, T. (1996) A Discovery Algorithm for Directed Cyclic Graphs. In *Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Conference* (F. Jensen and E. Horvitz, eds.), pp. 462-469, Morgan Kaufmann, San Francisco.
- Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7, 1393-1512.
- Robins, J., Scheines, R., Spirtes, P., and Wasserman, L. (forthcoming) "Uniform Consistency in Causal Inference", *Biometrika*.
- Rubin, D. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, **2** pp. 1-26.

- Sewell, W. and Shah, V. (1968). Social class, parental encouragement, and educational aspirations. *American Journal of Sociology*, **73**, pp. 559-572.
- Shafer, S. (1996). *The Art of Causal Conjecture*. Cambridge, MA, MIT Press.
- Shipley, W. (2000) *Cause and Correlation in Biology*. Cambridge University Press, Cambridge, England.
- Sosa, E., and Tooley, M. (eds.) (1993) *Causation*. Oxford University Press, New York.
- Spearman, C. (1904) General intelligence objectively determined and measured. *American Journal of Psychology* **15**, 201-293.
- Spirtes, P., Glymour, C., and Scheines, R. (2000) *Causation, Prediction, and Search*, MIT Press, Cambridge, MA.
- Spirtes, P., (1995) Directed Cyclic Graphical Representation of Feedback Models, in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. by Philippe Besnard and Steve Hanks, Morgan Kaufmann Publishers, Inc., San Mateo.
- Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*, Wiley, New York.
- Spirtes, P., and Richardson, T. (1996), A Polynomial Time Algorithm For Determining DAG Equivalence in the Presence of Latent Variables and Selection Bias, *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*.
- Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociological Methods and Research*, **27**, 148-181.
- Strotz, R., and Wold, H. (1960) Recursive versus nonrecursive systems: an attempt at synthesis. *Econometrica*, **28**, 417-427.
- Sullivan, J., & Feldman, S. (1979). *Multiple indicators: an introduction*. Sage Publications, Beverly Hills, CA.
- Swanson, N., and Granger, C. (1997) "Impulse Response Function Based on a Causal Approach to Residual Orthogonalization in Vector Autoregression". *Journal of the American Statistical Association* **92**(437), pp. 357-367.
- Verma, T. and Pearl, J. (1990b). Equivalence and synthesis of causal models in *Proceedings of the Sixth Conference on Uncertainty in AI*. Association for Uncertainty in AI, Inc., Mountain View, CA.
- Wright, S. (1934). The method of path coefficients, *Annals of Mathematical Statistics* **5**, 161-215.
- Zhang, J. (2002) *Consistency in Causal Inference Under a Variety of Assumptions*, Masters Thesis, Department of Philosophy, Carnegie Mellon University.

Errors in PC and FCI

Selection bias

Mags

Scales and patterns for measured variables

Search by hand; search by lisrel

Lauritzen on mixed

What about others? Not closed under marginalization or conditionalization