8-4-2009

A Tutorial On Causal Inference

Peter Spirtes
Carnegie Mellon University

A Tutorial On Causal Inference

Peter Spirtes

August 4, 2009

Technical Report No. CMU-PHIL-183

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

A Tutorial On Causal Inference

Peter Spirtes[1]

1. Introduction

The goal of many sciences is to understand the mechanisms by which variables came to take on the values they have (i.e. to find a generative model), and to predict what the values of those variables would be if the naturally occurring mechanisms in a population[2] were subject to outside manipulations. For example, a randomized experiment is one kind of manipulation, which substitutes the outcome of a randomizing device to set the value of a variable, such as whether or not a particular diet is used, instead of the naturally occurring mechanism that determines diet. In non-experimental settings, biologists gather data about the gene activation levels in normally operating systems, and seek to understand which genes affect the activation levels of which other genes, and seek to predict what the effects of intervening to turn some genes on or off would be; epidemiologists gather data about dietary habits and life expectancy in the general population and seek to find what dietary factors affect life expectancy and to predict the effects of advising people to change their diets; econometricians gather data on many econometric variables and seek to find what causes growth in GDP and inflation, and to predict the effects of policies a central bank could adopt. Finding answers to questions about the mechanisms by which variables come to take on values, or predicting the value of a variable after some other variable has been manipulated, is characteristic of causal inference. If only observational (non-experimental) data is available, predicting the effects of manipulations typically involves drawing samples from one density (of the unmanipulated population) and making inferences about the values of a variable in a population that has a different density (of the manipulation population).

The problem of causal inference is one that occurs in many domains, and there have been a wide variety of different causal models and inference algorithms that have been developed. Nevertheless, many of the basic problems and basic assumptions remain the

---

[1] I would like to thank Isabelle Guyon, Constantin Aliferis, and Greg Cooper for many helpful comments.
[2] Here, the "population" is simply a collection of instantiations of a set of random variables. For example, it could consist of a set of barometer readings and atmospheric pressures in different locations at a given

same across domains. In addition, although there are some superficial similarities between traditional supervised machine learning problems and causal inference (e.g. both employ model search and feature selection, the kinds of models employed overlap, some model scores can be used for both purposes), these similarities can mask some very important differences between the two kinds of problems.

The goal of this tutorial is twofold: to provide a description of some basic causal inference problems, models, algorithms, and assumptions in enough detail to understand recent developments in these areas; and to compare and contrast these with machine learning problems, models, algorithms, and assumptions.

1.1. History

Traditionally, there have been a number of different approaches to causal discovery. The gold standard of causal discovery has typically been to perform planned or randomized experiments [Fisher, 1971, #91994]. There are obvious practical and ethical considerations that limit the application of randomized experiments in many instances, particularly on human beings. Moreover, recent data collection techniques and causal inference problems raise several practical difficulties regarding the number of experiments that need to be performed in order to answer all of the outstanding questions (Eberhardt & Clark Glymour, 2006; Eberhardt, Glymour, & Scheines, 2005). First, there may be thousands of potential causes of a given effect that need to be investigated. For example, a given gene may have thousands of other genes that are potential regulators. Also, the number of genes for which it would be desirable to know which genes they were regulated by also numbers in the thousands. In addition, biologists would also like to understand the pathways by which genes regulate each other. For example, is gene X an indirect cause of gene Y (that is it regulates the activity level of gene Y via modifying the activity level of gene Z), or is it a direct cause of gene Y (that is, X modifies the behavior of Y not via modifying the activity level of other genes)? And if the former, which genes are the intermediate genes, and do they occur on the same pathway, or on parallel pathways? Questions about causal pathways might require manipulating multiple genes simultaneously (e.g. manipulating gene X to different values while holding gene Z

---

time, or the readings of a single barometer and atmospheric pressure over time, or a combination of these.

fixed to see whether *Z* is on a causal pathway from *X* to *Y*). Currently, it is not feasible to perform the number of experiments that would be required to answer all of these questions. Similar problems (on a smaller scale) exist in discovering cell signaling involving protein-protein interactions, where experimental manipulations have been combined with recently developed causal discovery algorithms (Sachs, Perez, Pe'er, Laurrenburger, & Nolan, 2005).

In the absence of experiments, search for causal models is often informal, and based on a combination of background knowledge together with statistical tests of the causal models, especially in the social sciences. If a model is rejected by a statistical test, the researcher looks for a modification of the original hypothesized model that will pass a statistical test. The search typically halts when a model that is compatible with background knowledge does not fail a statistical test. Often, the final model is presented, and the search itself is not described. Informal searches of this kind fail to account for multiple testing problems, can potentially lead to severe overfitting problems, and can only be as reliable as the reliability of the background knowledge, and to the extent to which the space of alternatives compatible with the background knowledge was searched. (For an example of a case where a search is described, see Rodgers & Maranto, 1989). Rodgers and Maranto show that different social science disciplines often start from very different causal models, and have different background "knowledge". Furthermore, unless the background knowledge is very extensive, or the number of variables is tiny, it is not feasible to estimate and test all of the models compatible with background knowledge. This is further complicated by the fact that for reliable causal inference it is not sufficient to find one model that passes a statistical test; instead it is necessary to find all such models. (See section 5.2.)

In the last several decades, new kinds of causal models have been proposed: causal Bayesian networks (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993), chain graphs (Lauritzen & Richardson, 2002), mixed ancestral graphs (Richardson & Spirtes, 2002b)) and many others. In addition, a host of new automated searches for causal inference have been developed (Spirtes et al., 1993; Heckerman, 1998; Chickering, 2003).

Causal inference is a complex process that can involve a great deal of pre-processing of the data (e.g. to correct for measurement errors of various kinds), many different kinds

of data (e.g. observational, experimental) and study designs (e.g. case-control, time series, meta-analyses). Results from one study may suggest further studies of the same or different kinds. There are also a wide variety of different kinds of models that can be given causal interpretations that are employed in the sciences, e.g. differential equation models, finite difference models, structural equation models (both linear and non-linear), graphical models (Bayesian networks of various kinds), HMMs (Hidden Markov Models), ICA models (Independent Component Analysis), Factor Analysis models, neural networks, etc. (Mitchell, 1997).

In order to highlight what is distinctive about causal inference, rather than describing the entire complex process of causal inference or surveying all of the kinds of causal models, this tutorial will examine in detail some very simplified examples of causal inference. The examples illustrate a number of distinctive features of causal inference as opposed to machine learning. The basic examples concern predicting the value of a target variable after a population has been subjected to some external intervention or manipulation from sample data that has been collected prior to the manipulation. For example, under normal operation, atmospheric pressure and barometer readings are highly associated, and either variable can be accurately predicted from the other. If the atmospheric pressure is manipulated somehow, then the new value of the atmospheric pressure still accurately predicts the barometer reading; however, if the barometer reading is manipulated (e.g. by spinning the dial) the new barometer reading does not accurately predict atmospheric pressure. There are several varieties of problems that will be discussed, but the fundamental question is when and how it is possible to tell from data collected prior to a manipulation (e.g. on the normal relation between the barometer reading and the atmospheric pressure) how to accurately predict a target variable after a manipulation. (The contrast between typical machine learning problems and typical causal inference problems is drawn more formally and in more detail in sections 3.2 and 7 respectively.) In some cases, such predictions cannot be reliably made at all from non-experimental data (thus the maxim "correlation is not causation"), but in other cases there are reliable inference methods. If the time order in which events occurred is known, and there are no hidden common causes, and no deterministic relationships among variables, then finding the causes of a variable X reduces to the problem of finding the smallest set

of predictors among the temporal predecessors of *X* that contained as much information about *X* as all of the predecessors of *X* do.[3] However, this simple algorithm

Section 2 provides an introduction to one kind of causal model and explains the concept of "manipulation" in more detail. Sections 3 and 4 give a statistical perspective on standard supervised machine learning problems. Sections 5, 6 and 7 contrast the standard supervised machine learning problem with an analogous causal inference problem, and describe the three basic steps in solving the causal inference problem. Section 8 describes the extra complications in causal inference introduced by the possibility of unmeasured common causes. Section 9 briefly describes some open problems in causal inference. Section 10 provides a brief summary, and section 11 is an appendix defining some statistical and graph theoretic terms.

2. Manipulating versus conditioning

A (parametric) *statistical model* (with free parameters) is a set of probability densities, that can be mapped into a single density by specifying the values of the free parameters (e.g. a family of multivariate Gaussian densities)[4]. For example, a Hidden Markov Model with a fixed structure but free parameters is a statistical model that represents a certain set of densities. A *causal model* (with free parameters) not only specifies a set of probability densities over a given set of variables, it also specifies the post-manipulation probability density over a given set of variables, for each manipulation that can be performed on the population. A causal model with free parameters is mapped into a single probability densities given the values of the free parameters *and* a specification of a manipulation.

For example, given an imperfect barometer, atmospheric pressures (plus some noise) causes barometer readings, but barometer readings do not cause atmospheric pressures. A causal model entails a joint density if the barometer readings are manipulated: if the

---

[3] Given the Causal Markov assumption described in section 5.1.1, *X* is independent of all temporal predecessors that are not direct causes conditional on its direct causes. This set is unique under the conditions given [Pearl, 2000, #15729]. This method for inferring causation is similar to "Granger causation" [Granger, 1969, #68098].

[4] In the nomenclature of machine learning, what this tutorial calls a "model (with free parameters)" is often called a "model family" or "learning machine" and a "model (with fixed parameter values)" is often called a "model instance" or "model".

barometer readings are set to *x* (by spinning the dial by force) then the probability density of the atmospheric pressures would remain the same and the probability of the barometer readings being *x* equals 1. Similarly, a causal model entails a different post-manipulation density if the atmospheric pressure is manipulated to *y*. (See section **Error! Reference source not found.** for more details.)

Often, a causal model is specified by giving a statistical model with free parameters, and a causal graph that describes the causal relations between variables. Together the statistical model and the causal graph entail a function of the free parameters and the manipulation that determine a unique (post-manipulation) probability density. This is described in more detail in section 2.4.

The most frequently used causal models belong to two broad families. One kind of model, including structural equation models (SEMs) specifies the *value of a variable* as a function of the values of its causes (typically including some unmeasured noise.) Causal Bayesian networks (and related models), specify a *probability density for a variable* as a function of the values of its causes. However, the two kinds of models are closely linked, as explained in section **Error! Reference source not found.**.

2.1. Structural Equation Models (SEMs)

Although this tutorial focuses, for didactic reasons, on a simple linear model, the setting purposely bears similarity with the classical machine learning setting. Estimating the value of a continuous variable in classical statistics is nothing but a regression problem (and if the variable is categorical it is a classification problem). One goal of this tutorial is to contrast regression with causal inference, which is a problem involving a specified change to the underlying structure of the data generating system that leads to a change in data density between model construction and model use. (There is an important distinction between the data generating "system" (the world) and the causal "model" (a theory about the world) used to make predictions, since the latter need not be identical or even from the same family as the former). In the example, the data are assumed to be generated by a <u>structural equation model</u>, and the predictive models also come from the same model family.

The statistical setup is a standard one. There is a population of units, where depending upon the problem, the units could be people, cities, cells, genes, etc. It is assumed that there is a probability distribution over the population, which assigns probabilities to each measurable subset (event) of the population.[5] Each unit also has a set of properties at a time, where the properties are represented by random variables, which are functions from the units to real numbers.[6] For example, the property of height at a time can be represented by a number representing height in inches, and if the height is constant over the temporal interval relevant to the problem, the time index can be suppressed. Although the suppression of the time index is not always possible, in the simplified cases considered here, the temporal indices of the random variables will be suppressed. The probability distribution over the random variables is determined by the probability distribution over the units in the population. I will assume that the joint probability distribution over the random variables can be represented by a probability density function [Bickel and Doksum, 2000, #90811].

The set of random variables in a structural equation model (SEM) can be divided into two subsets, the "error variables" or "error terms," and the substantive variables (for which there is no standard terminology in the literature). The substantive variables are the variables of interest, but they are not necessarily all observed. Which variables are substantive, and which variables are error terms can vary with the analysis of the problem. In SEM *K*, shown in

Figure 1, the substantive variables are the variables *B* (*Barometer Reading*), *A* (*Atmospheric Pressure*), and *R* (*Rainfall*). For each substantive variable such as *B* there is

[5] More formally, there is a probability space consisting of a population **S**, a sigma-field of events over **S**, and a probability function mapping members of the sigma-field into the real interval from 0 to 1 inclusive that obeys the Kolmorogov axioms of probability.

[6] Every random variable is a function f such that the inverse function f^{-1} is such that each Lebesgue measurable set of real numbers is mapped back to an event, i.e. a member of the sigma field.

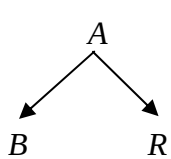
a corresponding error term for B that represents all of the causes of B that are not substantive variables, e.g. ε_B - for example, noise in the reading of the barometer. Each substantive variable such as B occurs on the left hand side of one equation that relates the value of B to the direct causes of B plus the error term ε_B on the right hand side of the equation, e.g. $= B = b_{B,A} \cdot A + \varepsilon_B$, where $b_{B,A}$ is a linear coefficient of A in the structural equation for B. (Bollen, 1989).

To simplify the examples, the error terms will (unrealistically) be assumed to be Gaussian and the structural equations will assumed to be linear, unless explicitly stated otherwise. In addition, the parameter values were chosen in order to simplify the calculations, rather than to be realistic.

SEMs have two forms - a free parameter form, and a fixed parameter form. These are illustrated in

Figure 1 and Figure 2 respectively (and the various terms in the figures are explained in the following paragraphs). In the free parameter form, the linear coefficients in the structural equations (e.g. $b_{B,A}$), and the covariance matrix of the error terms (e.g. $\sigma^2(\varepsilon_B)$) are variables. In the fixed parameter form, the linear coefficients in the structural equations (e.g. $b_{B,A}$), and the covariance matrix among the error terms (e.g. the variance $\sigma^2(\varepsilon_A)$) are constants. The context will make it clear whether $b_{B,A}$ refers to a variable (in a SEM with free parameters) or a constant (in a SEM with fixed parameters.)

Causal Graph



Entailed Covariance Matrix

	A	B	R
A	$\sigma^2(\varepsilon_A)$	$b_{B,A}\sigma^2(A)$	$b_{R,A}\sigma^2(A)$
B	$b_{B,A}\sigma^2(A)$	$b^2_{B,A}\,\sigma^2(A) + \sigma^2(\varepsilon_B)$	$b_{B,A} \cdot b_{R,A}\sigma^2(A)$
R	$b_{R,A}\sigma^2(A)$	$b_{B,A} \cdot b_{R,A}\sigma^2(A)$	$b^2_{R,A}\,\sigma^2(A) + \sigma^2(\varepsilon_R)$

Equations

$A = \varepsilon_A$

$B = b_{B,A} \cdot A + \varepsilon_B$

$R = b_{R,A} \cdot A + \varepsilon_R$

Example of Conditional Expected value

$E_{K(\theta)}(A|B{=}m) =$

$E(\varepsilon_A) + b_{B,A} \cdot \sigma^2(A)/\sigma^2(B) \cdot (m - b_{R,A} \cdot E(\varepsilon_A) + E(\varepsilon_B))$

Free Parameters

$\sigma^2(\varepsilon_A)$, $\sigma^2(\varepsilon_B)$, $\sigma^2(\varepsilon_R)$

$b_{R,A}$, $b_{B,A}$, $E(\varepsilon_A)$, $E(\varepsilon_B)$, $E(\varepsilon_R)$

Example of Total Effect

Total Effect of B on $A = 0$

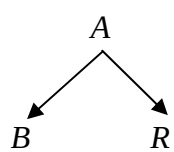
Example of Manipulated Expected value

$E_{K(\theta)}(A||B{=}m) = 0$

Figure 1: SEM *K* With Free Parameters

Each SEM is associated with a directed graph whose vertices include the substantive variables, and that represents both the causal structure of the model and the form of the structural equations. (Typically, the graph is selected prior to the form of the equations or the coefficients, from causal background knowledge, together with an informal or formal search. See sections 3.2.2, 7.1.1, 8.1.1 for more details about search.) There is a directed edge from A to B ($A \rightarrow B$) if the coefficient of A in the structural equation for B is non-zero. The coefficient $b_{B,A}$ of A in the structural equation for B is the *structural coefficient associated* with the edge $A \rightarrow B$. In general, the graph of a SEM may have cycles (i.e. directed paths from a variable to itself), and may explicitly include error terms with double-headed arrows between them to represent that the error terms are dependent (e.g. $\varepsilon_A \leftrightarrow \varepsilon_B$); if no such edge exists in the graph, the error terms are assumed to be independent. If a variable has no arrow directed into it, then it is <u>*exogenous*</u>; otherwise it is <u>*endogenous*</u>. In K, A is exogenous and B and R are endogenous. If the graph has no directed cycles and no double-headed arrows, then it is a <u>*directed acyclic graph*</u> (DAG).

Causal Graph



Entailed Covariance Matrix

	A	B	R
A	1	.6	.8
B	.6	1	.48
R	.8	.48	1

Equations

$A = \varepsilon_A$

$B = .6 \cdot A + \varepsilon_B$

$R = .8 \cdot A + \varepsilon_R$

Fixed Parameter Values

$\sigma^2(\varepsilon_A) = 1$, $\sigma^2(\varepsilon_B) = .64$,

$\sigma^2(\varepsilon_R) = .36$

$b_{R,A} = .8$, $b_{B,A} = .6$

$E(\varepsilon_A) = E(\varepsilon_B) = E(\varepsilon_R) = 0$

Example of Conditional Expected Value

$E_{K(\boldsymbol{\theta})}(A|B{=}m) = 0.6 \cdot m$

Example of Total Effect

Total Effect of B on $A = 0$

Example of Manipulated Expected Value

$E_{K(\boldsymbol{\theta})}(A||B{=}m) = 0$

Figure 2: SEM *K* With Fixed Parameters $\boldsymbol{\theta}$

Which variables are "substantive" depends upon the analysis, not upon the way the world is. Suppose for example, SEM K of Figure 1were modified so that A was not included among the substantive variables. In that case, the "other non-substantive causes" of B and R (i.e. both ε'_B and ε'_R) would both contain A, which would create a dependency between the two error terms. This dependency can be represented in the corresponding graph by introducing ε'_B and ε'_R into the graph as parents of B and R respectively, and introducing a double-headed arrow between them, i.e. $\varepsilon_B \leftrightarrow \varepsilon_R$. A correlation between error terms ε_B and ε_R is typically taken to be due to a common cause of B and R that is not included among the substantive variables, but the representation is less specific than $B \leftarrow A \rightarrow R$ because the former does not specify how many common causes there are or there relationships to each other, while the latter does (Simon, 1985; Bollen, 1989).

The free parameters of K are the *linear coefficients* corresponding to edges in the associated graph and the *variances of the error terms*. $K(\theta)$ represents the fixed parameter SEM where the free parameter have been assigned fixed values according to an assignment θ, e.g. $\theta = \{\sigma^2(\varepsilon_A) = 1, \sigma^2(\varepsilon_B) = .64, \sigma^2(\varepsilon_R) = .36, b_{B,A} = .6, b_{R,A} = .8, E(\varepsilon_A) = E(\varepsilon_B) = E(\varepsilon_R) = 0\}$.

It is assumed that the causal order in the graph is compatible with the time order of the variables and the system is in a stationary state when the variables are measured.

Let $\mathbf{X} = \{A, B, R\}$ be the set of all substantive variables in K.[7] The structural equations together with the probability density of the error terms in $K(\theta)$ entail a probability density $f_{K(\theta)}(\mathbf{X})$ over the substantive variables $\mathbf{X}$, as shown in

Figure 1.

In matrix form, the structural equations are $\mathbf{X} = \mathbf{BX} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the set of all error terms, and $\mathbf{B}$ is the structural linear coefficient matrix. If the covariance matrix among the $\boldsymbol{\varepsilon}$ is Σ_{ε}, $\Sigma_{K(\theta)} = (\mathbf{I}-\mathbf{B})^{-1}\Sigma_{\varepsilon}(\mathbf{I}-\mathbf{B})^{-1\mathrm{T}}$, where $\mathbf{I}$ is the identity matrix, $(\mathbf{I}-\mathbf{B})^{-1}$ is the inverse of $(\mathbf{I}-\mathbf{B})$, and $(\mathbf{I}-\mathbf{B})^{-1\mathrm{T}}$ is the transpose of the inverse of $(\mathbf{I}-\mathbf{B})$.

[7] Individual variables are in italics, and sets of variables are in boldface. Variables are capitalized, and values of variables are lower-case.

There is a simple rule for calculating cov(X_j, X_i) from the associated DAG. A *trek* between X_j and X_i is either a directed path from X_j to X_i, or a directed path from X_i to X_j, or a pair of directed paths from some variable X_k to X_j and X_i that intersect only at X_k. (A trek between X_j and X_i is also a trek between X_i and X_j.) The *source* of a trek is the unique vertex on the trek that has no edge directed into it. A *trek product* is the product of the linear structural coefficients associated with each edge on the trek times the variance of the source of the trek. For example, the trek product of $B \leftarrow A \rightarrow R$ is $b_{B,A} \cdot b_{R,A} \cdot \sigma^2(A)$. The covariance $\mathrm{cov}_{K(\boldsymbol{\theta})}(B, R)$ is the sum of all trek products between B and R (Spirtes, Glymour, & Scheines, 2001). If there are no treks between two variables, then the covariance between the two variables is zero. SEM K has one trek between B and R, so $\mathrm{cov}_{K(\boldsymbol{\theta})}(B,R) = b_{B,A} \cdot b_{R,A} \cdot \sigma^2(A)$.

The Gaussian density $f_{K(\theta)}(\mathbf{X})$ is completely characterized by the covariance matrix $\Sigma_{K(\theta)}$ over the substantive variables, and the expected values μ of the substantive variables. $f_{K(\theta)}(\mathbf{X}) \sim \mathrm{N}(\Sigma_{K(\theta)}, \mu)$ is a function of the fixed parameter values, e.g. $\mathrm{cov}_{K(\theta)}(B,R) = b_{B,A} \cdot b_{R,A,} \cdot \sigma^2(A)$. If, for a given covariance matrix Σ there exists an assignment of parameter values θ of K so that the entailed covariance matrix is Σ (i.e. $\Sigma_{K(\theta)} = \Sigma$) K is said to *represent* Σ. For example, the covariance matrix in Figure 2 is represented by K, because the assignment θ of values to the free parameters entails that covariance matrix among the substantive variables. If for each Σ that can be represented by K, there is a unique assignment of values to the free parameters θ such that $\Sigma_{K(\theta)} = \Sigma$, then the free parameters of K are *identifiable*. The free parameters of any linear SEM whose graph is a DAG (with no unmeasured substantive variables) are identifiable. The maximum likelihood estimates of the linear coefficients relating a variable X to its parents in a DAG are simply the coefficients obtained by regressing X on its parents (Bollen, 1989). If the free parameters of a SEM are identifiable, then given a covariance matrix Σ that is represented by K, the values of the free parameters are uniquely determined by Σ.

For purposes of illustration, it is assumed that the graphs are DAGs unless explicitly stated otherwise. The assumptions of Gaussian error terms, linear structural equations,

and DAGs simplify the examples while still illustrating the basic problems for causal inference and the basic strategies for solving the problems.

Given the independent error terms in SEM K, for each θ, SEM K entails both a set of conditional independence relations among the substantive variables, and that the joint density over the substantive variables *factors according to the graph*, i.e. the joint density can be expressed as the product of the density of each variable conditional on its parents in the graph. For example, $f_{K(\theta)}(A,B,R) = f_{K(\theta)}(A)f_{K(\theta)}(B|A)f_{K(\theta)}(R|A)$ for all θ. This factorization in turn is equivalent to a set of conditional independence relations among the substantive variables (Lauritzen, Dawid, Larsen, & Leimer, 1990).

$I_f(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ denotes that **X** is independent of **Y** conditional on **Z** in density f, i.e. $f(\mathbf{X}|\mathbf{Y},\mathbf{Z}) = f(\mathbf{X}|\mathbf{Z})$ for all $f(\mathbf{X}|\mathbf{Z}) \neq 0$. (In cases where it does not create any ambiguity, the subscript f will be dropped). $I_{M(\theta)}(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ denotes that **X** is independent of **Y** conditional on **Z** in the density over the substantive variables entailed by $M(\theta)$. $I_M(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ denotes that a SEM M (with free parameters) entails $I_{M(\theta)}(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ for *all* assignment of values θ to its free parameters; in other words **X** is independent of **Y** conditional on **Z** in every density represented by M. If $I_M(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ then M is said to *entail* that **X** and **Y** are independent conditional on **Z**. Since $I_M(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ does not depend upon the parameter values of M, it is possible to determine whether $I_M(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ from the graph of M using the graphical criterion, "d-separation", which is defined in the following paragraph, after a series of preliminary definitions.

A variable B is a <u>*collider on a path*</u> U if and only if U contains a subpath $A \rightarrow B \leftarrow C$. For disjoint sets of vertices **X**, **Y**, and **Z** in a DAG G, **X** is <u>*d-connected*</u> to **Y** given **Z** if and only if there is an acyclic path U between some member X of **X**, and some member Y of **Y**, such that every collider on U is an ancestor of **Z**, and every non-collider on U is not in **Z**.[8] For disjoint sets of vertices, **X**, **Y**, and **Z**, **X** is <u>*d-separated*</u> from **Y** given **Z** if and only if **X** is not d-connected to **Y** given **Z**. **X** is d-separated from **Y** conditional on **Z** in DAG G if and only if $I_{G(\theta)}(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ in every set of parameter values θ in which the error terms are independent (Pearl, 1988).

---

[8] For both the d-separation relation and the independence relation, if **X** contains a single vertex X, X will be written instead of $\{X\}$, and similarly for **Y** and **Z**. D-connection can also be defined for cyclic graphs and graphs with double-headed arrows (Spirtes, 1995; Koster, 1999; Cox & Wermuth, 1996b).

For example, in SEM K, B and R are d-connected conditional on the empty set by the path $B \leftarrow A \rightarrow R$. However, B and R are d-separated conditional on A because the only path between B and R ($B \leftarrow A \rightarrow R$) has a non-collider A that is in the conditioning set. Hence for every density f represented by K (i.e. for all values of the free parameters), $I_f(B,R|A)$.

2.2. An Aside on SEMs and Bayesian Networks

A <u>Bayesian network</u> is a pair of $<G, P>$, where G is a directed acyclic graph and a P is a probability density such that if **X** and **Y** are d-separated conditional on **Z** in G, then **X** and **Y** are independent conditional on **Z** in G. If the error terms in a SEM with a DAG G are jointly independent, and $f(\mathbf{V})$ is the entailed density over the substantive variables, then $<G,f(\mathbf{V})>$ is a Bayesian network.

A SEM treats each unit in the population as a deterministic system, but different units in the population differ in the values of the error terms, a Bayesian network does not require that each unit be a deterministic system (although it could be). It simply constrains the joint population density. It is always possible to turn a Bayesian network into a SEM by simply defining the error terms to be the

1. Bayes - local, global, or factorization? cyclic?
2. do all Bayesian networks have SeM representation?
3. do all SEMs have Bayesian network representation?

It is possible that **X** and **Y** are d-connected conditional on **Z** in some SEM G, but that $I_{G(\theta)}(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ for *some* (but not all) θ. An example of this is given in section 5.1.2. (There are no examples of this in SEM K).

In the case of Gaussian density f, **X** and **Y** are independent conditional on **Z** if and only if each variable $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, $\mathrm{cov}_f(X,Y|\mathbf{Z}) = 0$ (where $\mathrm{cov}_f(X,Y|\mathbf{Z})$ is the covariance between X and Y conditional on **Z** in density f). For Gaussian densities , the conditional covariance does not depend upon the particular value of **Z** conditioned on (although the conditional expected value does).

2.3. Conditioning

The density of *A conditional* on $B = m$ (denoted $f(A|B=m)$) represents the density of A in the *subpopulation* where $B = m$, and is defined from the joint density as $f(A|B=m) = f(A,B=m)/f(B=m)$. The conditional density depends only upon the joint density (assuming the values of the variables conditioned on do not have density 0), and does not depend upon the causal relationships. When a variable B is conditioned on, this intuitively represents *seeing* the probability density for the subpopulation in which B has the value conditioned on.

Conditional probabilities are useful for problems (e.g. diagnosis) in which the value of the variable of interest is expensive or difficult to measure, but other related variables that have information about the variable of interest are more easily observed. For example, the barometer reading, even if noisy, can be easily read than finding the exact value of the atmospheric pressure. Similarly, it is often easier to diagnose a disease from symptoms, than to directly test for the presence of some virus, or other condition.

So far the SEM has been given a purely statistical interpretation, i.e. it represents a set of probability densities. The next section describes how to give SEMs causal interpretations, and how to use them to calculate the effects of manipulations.

2.4. Manipulating

In contrast to conditioning, a post-manipulation probability density is not a density in a subpopulation of an existing population, but is a density in a (possibly) hypothetical population formed by manipulating the properties in a causal system.[9]

The causal interpretation is an extension of the standard statistical interpretation that assumes that each element of a SEM not only describes the pre-manipulation random variables, but also describes the post-manipulation random variables . There are a number of different causal interpretations that extend the standard statistical interpretation, which disagree on such issues as whether there is genuine indeterminism in the world, and the role of counterfactuals in causal reasoning [Dawid, 2007, #52556]. Rather than discuss

[9] "Manipulation" is a causal concept. No reduction of causal concepts to non-causal concepts will be attempted here, and so no definition of manipulation will be given. Instead, some axioms postulating connections between causal models and densities will be given (section 5.1).

all of these issues, I will present one simple causal interpretation here, which assumes that each causal system is deterministic, but may appear indeterministic because error terms are not observed.

For example, in an experiment in which B is randomized (which might be impossible to do in practice), whatever causes of B existed in the existing population are replaced by the outcome of a randomization device as the sole cause of B. If the barometer reading is randomized, then the dial reading is set by spinning the dial on the basis of the outcome of some randomizing device, and the causal connection between the atmospheric pressure and the barometer reading is severed. In contrast to conditioning which corresponds to *seeing*, manipulating corresponds to *doing*; the operation of manipulation is also sometimes referred to as a "do-operator" (Pearl, 2000). The assumption is made that the only direct effect of the manipulation is on the variable being manipulated - any other effects are due to the change in the variable being manipulated. Hence, in SEM K, a manipulation of the barometer reading is not carried out by forcing a change on the atmospheric pressure - that would be a manipulation of the atmospheric pressure.

Intuitively manipulating is different than conditioning. Given the normal, unmanipulated operation of the barometer, in the subpopulation in which the barometer reading is m, there is a narrow density of atmospheric pressures centered around m - the density of atmospheric pressures is not the same as the density of atmospheric pressures in the subpopulation where the barometer reading is 0. In contrast, when the barometer reading is manipulated to m by spinning the dial, the density of atmospheric pressures is the same as if the dial is not manipulated. In this case, the density of atmospheric pressures after the barometer reading is manipulated does not equal the density of atmospheric pressures in the subpopulation where a normally operating barometer has a reading of m.

More formally, corresponding to each random variable B and a manipulation of a variable A to the value a, is a subjunctive random variable $B_{A=a}$, which is the value B would have if A were manipulated to have the value a [Rubin, 1974, #17334]. In cases, where it does not lead to any confusion, I will simply refer to the subjunctive random

variable $B_{A=a}$ as the post-manipulation value of B, and B itself as the pre-manipulation value of B.[10]

The manipulation of A to a is assumed to always be successful, so $A_{A=a} = a$. Under the causal interpretation, it is assumed not only that the random variables satisfy the structural equations, but also that the subjunctive values satisfy the manipulated structural equations, for any manipulation. (In practice the equations are only approximately satisfied, and only for a limited class of manipulations.) In other words if $B = .6 \cdot A + \varepsilon_B$ and $A = \varepsilon_A$ for each unit in the population, and A is manipulated to a, then $B = .6 \cdot a + \varepsilon_B$ and $A_{a=a} = a$, or more colloquially the post-manipulation values of A and B also satisfy the set of structural equations, where the equation for A has been replaced by the equation $A = a$. The distribution over the error terms, and the manipulated structural equations determine a manipulated joint distribution over the post-manipulation variables.

In some cases, the density that results from manipulating is the same as the density that results from conditioning. Suppose that the atmospheric pressure is manipulated. Intuitively, when the atmospheric pressure is manipulated, the barometer reading will change in response to the change in atmospheric pressure. As explained below, SEM K entails that the density of barometer readings when atmospheric pressure is manipulated to a is the same as the density of barometer readings in the subpopulation where atmospheric pressure is equal to a. Manipulating the atmospheric pressure still leaves the mechanism relating atmospheric pressure to barometer readings intact.

A manipulation of a variable X_i in a population can be described by any of the following three kinds of equations of increasing generality:

1. $X_i = c$
2. $X_i = \varepsilon'_i$
3. $X_i = \sum_{X_j \in \mathbf{Parents'}(X_i)} b'_{i,j} X_j + \varepsilon'_i$,

where all of the variables are the post-manipulation variables, **Parents'**(X_i) is a new set of causes of X (which are included in the set of non-effects of X_i in the unmanipulated

[10] In full detail, there should be time indices associated with both A, and $B_{A=a}$, which I have omitted here for simplicity. I will assume that after the manipulation each unit reaches a stationary state, and that the post-manipulation value of B is the subjunctive value of B in the stationary state.

population). The first kind of replacement sets X_i to the same value for everyone in the population (e.g. no one is allowed to smoke at all). The second kind of replacement assigns a distribution to X_i that does not depend upon any causes X_i has in the unmanipulated population (e.g., people are assigned whether to smoke or not based on the flip of a coin, as in a randomized experiment.) The third kind of replacement changes the relationship between an effect and its non-descendants in the unmanipulated population, e.g. people are assigned whether to smoke or not based upon both the flip of a coin and their age (Strotz & Wold, 1960; Spirtes et al., 2001; Pearl, 2000). If no temporal indices are placed on the variables, it is assumed that the post-manipulation density reaches an equilibrium (at least temporarily until it is subject to another external shock). This can be problematic for causal models with cyclic graphs, which introduce an extra set of problems over and above those introduced by causal inference with DAGs (Fisher, 1970). If the error terms are continuously changing, then temporal indices for the variables should be introduced into the SEM.

In a causal model such as SEM $K(\theta)$, the post-manipulation population is represented in the following way, as shown in Figure 3. When B is manipulated, a new model $K(\omega)$ is created out of $K(\theta)$, where $\omega = \theta$, except that the free parameters corresponding to the manipulation are changed, thus changing the structural equations. For example, if in $K(\theta)$, B is manipulated to have an expected value of 5 and does not depend upon the value of A (a manipulation of type 2), then in $K(\omega)$, $E(\varepsilon'_B) = 5$, $b_{B,A} = 0$, and $B = \varepsilon'_B$. The new structural equations and density over the error terms entail a new joint density over the substantive variables, where $E_{K(\omega)}(B) = 5$. The causal graph for $K(\omega)$ is the same as that for $K(\theta)$ except that the edge from A to B has been removed, to reflect the new structural equations. The result of modifying the set of structural equations in this way can lead to a density in the randomized population that is not necessarily the same as the density in any subpopulation of the general population. (For more details see Spirtes et al., 2001; Pearl, 2000). See Figure 3 for examples of manipulations to SEM K.

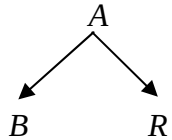
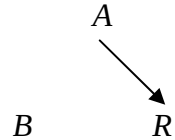
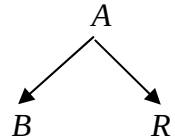
Causal Graph	**Causal Graph**	**Causal Graph**
		
Equations	**Equations**	**Equations**
$A = \varepsilon_A$	$A = \varepsilon_A$	$A = \varepsilon'_A$
$B = b_{B,A} \cdot A + \varepsilon_B$	$B = \varepsilon'_B$	$B = b_{B,A} \cdot A + \varepsilon_B$
$R = b_{R,A} \cdot A + \varepsilon_R$	$R = b_{R,A} \cdot A + \varepsilon_R$	$R = b_{R,A} \cdot A + \varepsilon_R$

Figure 3:(i) Unmanipulated; (ii) *B* Manipulated to 5; (iii) *A* Manipulated to 5

In a given population P with density $f(\mathbf{V})$, the density of $\mathbf{V}$ after a manipulation of X_i to a new density $f'(X_i|\mathbf{Parents}(X_i,G'))$ where G' is the causal graph after the manipulation, will be denoted as $f_P(\mathbf{V}||f'(X_i|\mathbf{Parents}(X_i,G')))$[11], where the double bar "||" notation denotes *manipulation* of X_i to the density $f'(X_i|\mathbf{Parents}(X_i,G')$.[12] When X_i is manipulated to a constant c, then the density of $\mathbf{V}$ after the manipulation is denoted as $f_P(\mathbf{V}||X_i=c)$. It is also possible to condition on the values variables take on after the manipulation: so the density of $\mathbf{X}$ in the subpopulation where $\mathbf{Y}=\mathbf{y}$ *after* manipulating X_i to a new density $f'(X_i|\mathbf{Parents}(X_i,G'))$ is denoted as $f_P(\mathbf{X}|\mathbf{Y}=\mathbf{y}||f'(X_i|\mathbf{Parents}(X_i,G'))$. (This is not necessarily the same as the post-manipulation density of $\mathbf{X}$ in the subpopulation which had $\mathbf{Y}=\mathbf{y}$ *prior* to the manipulation, because different units in the population might have $\mathbf{Y}=\mathbf{y}$ after the manipulation than had $\mathbf{Y}=\mathbf{y}$ before the manipulation.) For example, in a population P, if B is manipulated to a constant m, and then the resulting density of A is conditioned on the subset of units that have $R = r$ after the manipulation, the result is denoted as $f_P(A|R=r||B=m)$.

The post-manipulation density entailed by $K(\theta)$ is denoted $f_{K(\theta)}(\mathbf{X}|\mathbf{Y}=\mathbf{y}||f'(X_i|\mathbf{Parents}(X_i,G'))$. Figure 4 shows that SEM $L(\theta')$ has the same entailed density as SEM $K(\theta)$. However, the post-manipulation density $f_{K(\theta)}(A|R=r||B=m) \neq f_{L(\theta')}(A|R=r||B=m)$, even though $f_{K(\theta)}(A|R=r) = f_{L(\theta)}(A|R=r)$. Saying that $K(\theta)$ is the true

[11] There could also be a set of manipulations after the "||", if multiple variables are manipulated.
[12] The population is introduced in addition to the joint density because the results of the manipulation depend upon more than just the joint density. There are a number of different notations for post-

causal model for population P entails that $f_P(A|R{=}r||B{=}m) = f_{K(\theta)}(A|R{=}r||B{=}m)$, and similarly for all other manipulations.

In general, there is a parallel terminology for causal relations in the population based on the results of manipulations of the population, and for the entailed causal relations in a SEM based upon the representation of the manipulation in the SEM. For example, A can be said to be a cause of B in the population, and $K(\theta)$ can be said to entail that A is a cause of B.

2.4.1 Total Effect

In populations correctly described by linear SEMs, the change in X_i that results per unit change in manipulating X_j is the *total effect* of X_j on X_i. Intuitively, in the population, the total effect of A on B is non-zero, and the total effect of B on A is zero. If the total effect of X_j on X_i is non-zero, then X_j is a *cause* of X_i. In the barometer reading example, A is a cause of B, but B is not a cause of A.

The total effect of one variable on another can be due to several different mechanisms (or in a causal model, due to several different directed paths between two variables). For example, birth control pills affect embolisms both through preventing pregnancy (a negative effect) and through promoting blood clotting (a positive effect). The total effect entailed by a SEM can be calculated in the following way. In a SEM $K(\theta)$, the entailed total effect of X_j on X_i is equal to the sum over all directed paths from X_j to X_i of the product of the structural coefficients associated with each edge along a path (i.e. the *path sum*). In $K(\theta)$, the entailed total effect of A on R is 0.8, because there is one directed path from A to R ($A \rightarrow R$) and the product of coefficients of edges in that path is 0.8. The entailed total effect of R on A is zero, because there is no directed path from R to A; in other words, according to $K(\theta)$ manipulating the atmospheric pressure affects the barometer reading, but manipulating the barometer reading does not affect the atmospheric pressure.

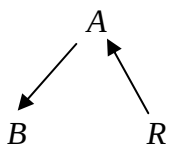
It is also possible to perform multiple manipulations simultaneously, by replacing the structural equations for multiple variables. Intuitively, in a population correctly described by a linear SEM, the *direct effect of X_j on X_i* relative to a set of variables **S** measures the

---

manipulation densities. The one used here is adapted from (Lauritzen, 2001)

strength of the effect of X_j on X_i that is not due to changing the values of any intermediate variables on a causal pathway from X_j to X_i; for example, it would measure the strength of the effect of the activity level of gene X on the activity level of gene Y if the effect of the activity level of X on the activity level of Y due to X affecting the activity levels of intermediate genes Z which in turn affected the activity level of Y were removed. In a linear SEM, the strength of an edge corresponds to the following manipulations: 1) manipulate all variables in **S** except X_i to some arbitrary fixed value c, and determine the manipulated value of X_i; 2) manipulate all variables in **S** except X_i and X_j to c, and manipulate X_j to (c+1), determining the new manipulated value of X_i; 3) take the difference between the two manipulated values of X_i. The direct effect is relative to a set of variables, because the set of variables determines which variables are manipulated. (For a discussion of total and direct effects see Bollen, 1989).

Causal Graph



Fixed Parameter Values

$\mathrm{var}(\varepsilon'_A) = .36$, $\mathrm{var}(\varepsilon'_B) = .64$, $\mathrm{var}(\varepsilon'_R) = 1$

$b_{A,R} = .8$, $b_{B,A} = .6$

$E(\varepsilon'_A) = E(\varepsilon'_B) = E(\varepsilon'_R) = 0$

Equations

$A = \varepsilon'_A$

$B = .6 \cdot A + \varepsilon'_B$

$R = .8 \cdot A + \varepsilon'_R$

Figure 4: SEM *L* With Fixed Parameters *θ'*

In a linear SEM $L(\theta')$ the entailed direct effect of R on A relative to the entire set of substantive variables in the SEM is simply the coefficient of the edge from R to A if there is one, and zero otherwise. For example in $L(\theta')$ in Figure 4, the direct effect of R on B is zero relative to **S'** = $\{A, B, R\}$. In step 1, R and A are manipulated to 5, and $L(\theta')$ entails that the manipulated value of B is $0.8 \cdot 5 + \varepsilon'_R = 4 + \varepsilon'_R$. In step 2, A is manipulated to 5, and R is manipulated to 6, and $L(\theta')$ entails that the manipulated value of B is $0.8 \cdot 5 + \varepsilon'_R = 4 + \varepsilon'_R$. In step 3, the direct effect of R on B relative to **S'** is $4 + \varepsilon'_R - (4 + \varepsilon'_R) = 0$. (In a population correctly described by $L(\theta')$, if both A and R are manipulated, then only the manipulation of A matters for the post-manipulation density of B.)

In contrast, suppose that the population is correctly described by $L(\theta')$, but **S** $=\{B, R\}$. R is shown to have a non-zero direct effect on B in $L(\theta')$ relative to **S** with the following

steps. In step 1, R is manipulated to some arbitrary value, say 5, and $L(\theta')$ entails that the manipulated value of B is to be $0.8 \cdot 0.6 \cdot 5 + \varepsilon'_R = 2.4 + \varepsilon'_R$ (because the product of coefficients in the path from R to B is $0.8 \cdot 0.6$). In step 2, R is manipulated to 6, and $L(\theta')$ entails that the manipulated value of B is $0.8 \cdot 0.6 \cdot 6 + \varepsilon'_R = 2.88 + \varepsilon'_R$. In step 3, the direct effect of R on B relative to **S** is $2.88 - 2.4 = .48$.

In a population correctly described by a linear SEM, if the direct effect of X_j on X_i relative to **S** is non-zero, then X_j is a *direct cause* of X_i relative to **S**. If SEM K is true, A is a direct cause of B and R relative to $\{A, B, R\}$. If SEM $L(\theta')$ is true, R is a direct cause of B relative to $\{B, R\}$, but R is not a direct cause of B relative to $\{A, B, R\}$.

A set **S** of variables is *causally sufficient* if every variable H that is a direct cause (relative to $\mathbf{S} \cup \{H\}$) of any pair of variables in **S** is also in **S**. Intuitively, a set of variables **S** is causally sufficient if no common direct causes (relative to **S**) have been left out of **S**. If SEM K is true then $\{A, B, R\}$ is causally sufficient, but $\{B, R\}$ is not because A is a common direct cause of B and R relative to $\{A, B, R\}$ but is not in $\{B, R\}$. If SEM C of Figure 5 is true, then $\mathbf{X} = \{X_1, X_2, X_3, X_4, H\}$ is causally sufficient. On the other hand, $\mathbf{X}\backslash\{H\}$ is not causally sufficient because H is a direct cause of $\{X_1, X_2, X_3, X_4\}$ relative to **X**, but H is not in $\mathbf{X}\backslash\{H\}$. If the observed set of variables is not causally sufficient, then the causal model is said to contain *unobserved common causes*, *hidden common causes*, or *latent variables*.

There are alternative representations of manipulations in models other than SEMs, including causal Bayesian Networks, often used for discrete variables (Spirtes et al., 2001), and potential response variables (Robins, 1986; Rubin, 1974), widely used in epidemiology.

SEM C SEM D

Figure 5: SEMs *C* and *D*

2.5. Causal Models versus Statistical Models

The causal interpretation of a SEM is an extension of the statistical interpretation of a SEM, which basically extends each element of the SEM from a description of the pre-manipulation random variables to a description of both the pre-manipulation *and* post-manipulation variables.

Under the statistical interpretation a SEM (with free parameters) represents a set of population densities of pre-manipulation variables; if the causal interpretation is added, the SEM also represents a set of population densities of post-manipulation variables. Considered purely as statistical models, $K(\theta)$ of Figure 2 (with free parameters) and $L(\theta')$ of Figure 4 are simply different parameterizations of the same statistical model, since they represent the same set of population densities of pre-manipulation variables (see section 5.2.2). On the other hand, $K(\theta)$ and $L(\theta')$ are different causal models, because they disagree on the populations densities of the post-manipulation variables.

Under the statistical interpretation, the graph of a SEM represents a set of entailed conditional independence relations (given by d-separation); if the causal interpretation is added, the edges of the graph also represent direct causal relations, and the manipulated graph also represents a set of entailed conditional independencies among the post-manipulation variables. Under the statistical interpretation of a SEM, for each unit in the population, the equations are satisfied by the values of the variables; if the causal interpretation is added, the equations are structural, i.e. the (manipulated) equations are also satisfied by the post-manipulation variables, and the equations describe the process that generated the data. Under the statistical interpretation, the linear coefficients in the equation for Y are the partial regression coefficients when Y is regressed on its parents in the graph; if the causal interpretation is added, then the linear coefficients describe the direct effect of each parent on Y.

Under the statistical interpretation, the error term of Y is a residual, i.e. it represents the difference between the predicted value of Y and the actual value of Y. If the causal interpretation is added, the error terms of Y also represents all causes of Y that are not substantive. For example, in $K(\theta)$, $R = 0.6 \cdot A + \varepsilon_R$, where $\sigma^2(\varepsilon_R) = 0.64$. In $K(\theta)$, 0.6 is the regression coefficient when R is regressed on A, and ε_R is the residual, i.e. the difference between the actual value of R, and the value of R predicted by the regression

on A. Under the causal interpretation, given that $K(\theta)$ is the true causal model, ε_R also represents all of the non-substantive causes of R. $L(\theta')$ can be interpreted as the same statistical model as $K(\theta)$ (but a different causal model.) In order to represent the same covariance matrix as $K(\theta)$, in $L(\theta')$, $R = \varepsilon'_R$, and $\sigma^2(\varepsilon'_R) = 1$. In $L(\theta')$, ε'_R can still be interpreted as a residual, i.e. the difference between the actual value of R, and the value of R predicted by the regression on its parents (i.e. the empty set, which amounts to using the expected value of R to predict R). However, if $K(\theta)$ is the true causal model, ε'_R cannot be interpreted as representing all of the non-substantive causes of A, because it has a different variance than ε_R, which represents all of the non-substantive causes of R.

3. Predicting Values from a Population Density

In order to explain the difference between causal inference and the usual machine learning problems, a number of different machine learning and causal problems will be described.

3.1. Inference Given the Population Density

Suppose the goal is to find a "good" predictor of the value of some target variable Y from the values of the observed covariates $\mathbf{O}$ for that unit. Ultimately, the prediction of the value of Y is performed by some prediction function $\hat{Y}(\mathbf{O})$. One traditional measure of how good the predictor $\hat{Y}(\mathbf{O})$ is in predicting Y is the mean squared prediction error (MSPE), which is equal to $E((Y - \hat{Y}(\mathbf{O}))^2)$, where the expected value is taken with respect to the density $f(\mathbf{O},Y)$ (Bickel & Doksum, 2000).[13]

In machine learning, there are many different kinds of prediction functions. For example, a parameterized neural network can serve as a prediction function. Typically, the prediction function is constructed from independent and identically distributed (i.i.d.) sample data from $f(\mathbf{O},Y)$ (so in machine learning terminology this is a supervised learning problem). For example, both the structure (the number of hidden nodes) and the

[13] Other measures of prediction error, such as the absolute value of prediction error or optimizing certain decision problems could be used, but would not substantially change the general approach taken here.

parameters (the edge strengths) of a neural network may be constructed from i.i.d. sample data.

However, in order to contrast standard supervised machine learning problems (described in this section) with causal learning problems (described in the next 3 sections) it is useful to start with the unrealistic, but simpler Problem 1 (learning from a given population density), before considering the more realistic Problem 2 (learning from a sample).

In all the problems that follow, $\mathbf{O} = \{O_1, O_2, \ldots O_m\}$ is a set of random variables used to predict the value of Y. $\mathbf{O}$ will be called the "predictor variables", and Y the "target variable". $f(\mathbf{O},Y)$ is the joint density of the predictor variables and the target variable.

Problem 1: Population predictive modeling

Input: $f(\mathbf{O},Y)$ and a target variable Y to be predicted for the unit.

Output: $\hat{Y}(\mathbf{O})$, an estimate of Y that minimizes the mean square prediction error (MSPE) of Y from the observed values $\mathbf{O}$.

Under weak regularity conditions, for $f(\mathbf{O},Y)$ the predicted value of Y that minimizes MSPE given $\mathbf{O}$ is $E_f(Y|\mathbf{O})$.

In the case of Gaussian variables, the population density f is characterized by the covariance matrix and the expected values. For a partitioned matrix of random variables $\mathbf{X_a}$, $\mathbf{X_b}$, $\mathrm{cov}(\mathbf{X_a},\mathbf{X_b}) = E_f((\mathbf{X_a} - E_f(\mathbf{X_a}))(\mathbf{X_b} - E_f(\mathbf{X_b})))$, $\mathrm{var}(\mathbf{X_a}) = \mathrm{cov}(\mathbf{X_a},\mathbf{X_a})$ and

$$\mathrm{var}(\mathbf{X_a},\mathbf{X_b}) = \begin{pmatrix} \mathrm{var}(\mathbf{X_a}) & \mathrm{cov}(\mathbf{X_a},\mathbf{X_b}) \\ \mathrm{cov}(\mathbf{X_b},\mathbf{X_a}) & \mathrm{var}(\mathbf{X_b}) \end{pmatrix}$$

The *linear* predictor that minimizes the MSPE for Y from $\mathbf{O}$ is

Equation 3-1

$$\hat{Y}(\mathbf{O}=\mathbf{o}) = E(Y \mid \mathbf{O}=\mathbf{o}) = E(Y) + \mathrm{cov}(Y,\mathbf{O})\mathrm{var}(\mathbf{O})^{-1} \cdot (\mathbf{o} - E(\mathbf{O})) = E(Y) + r_{Y.\mathbf{O}} \cdot (\mathbf{o} - E(\mathbf{O}))$$

where $\hat{Y}(\mathbf{O}=\mathbf{o})$ is the value of $\hat{Y}(\mathbf{O})$ for $\mathbf{O} = \mathbf{o}$, $r_{Y,Oj\cdot\mathbf{O}}$ is called the partial regression coefficient of O_j (i.e. the linear coefficient of O_j in the equation for Y), $r_{Y\cdot\mathbf{O}}$ is a vector of partial regression coefficients $r_{Y,Oj\cdot\mathbf{O}}$, and $\mathbf{o}$ is a vector of the values of members of $\mathbf{O}$. Equation 3-1 is the equation for Y regressed on $\mathbf{O}$. In the special case that Y is regressed

on a single variable O whose variance is 1, the partial regression coefficient of O is cov(Y,O).

For a Gaussian density, the regression equation is a predictor that minimizes the MSPE of Y from **O**, regardless of what the true causal model is. In the case where one variable is being conditioned on, and the variances of the variables are 1,

$$\hat{Y}(X = x) = E_f(Y \mid X = x) = \mathrm{cov}(Y, X) \cdot x = r_{Y,X.\mathbf{O}} \cdot x$$

For example, suppose the task is to predict rainfall from given values of the barometer reading (e.g. 4), and the atmospheric pressure (e.g. 5). This can be done by using the population covariance matrix to regress rainfall on the barometer reading and the atmospheric pressure. Using the value of the free parameters and entailed covariance matrix of $K(\theta)$ in Figure 2, the result is an equation for the regression of Y on $\{A, B\}$:

$$\hat{R}(A = 5, B = 4) = E(R \mid A = 5, B = 4) = \mathrm{cov}(R, \{A, B\})\mathrm{var}(\{A, B\})^{-1} \cdot \begin{bmatrix} 5 \\ 4 \end{bmatrix} =$$

$$\begin{bmatrix} .8 & .48 \end{bmatrix} \begin{pmatrix} 1 & .6 \\ .6 & 1 \end{pmatrix}^{-1} \begin{bmatrix} 5 \\ 4 \end{bmatrix} = \begin{bmatrix} .8 & 0 \end{bmatrix} \begin{bmatrix} 5 \\ 4 \end{bmatrix} = .8 \cdot 5 + 0 \cdot 4 = 4$$

where $r_{R,A\{A,B\}} = .8$ and where $r_{R,B\{A,B\}} = 0$.

Given the population covariance matrix, there is no subset of variables **O'** $\subset$ **O** such that regressing Y on **O'** produces a lower MSPE than regressing on **O**. Intuitively, this is plausible because in the population density $I_f(R,B|A)$, that is the reading of the barometer contains no information about the rainfall once the true atmospheric pressure is known. There is no point in introducing hidden variables or doing any kind of model search in order to improve the prediction, because the regression equation (using the *population* covariance matrix) cannot be improved on in terms of minimizing the MSPE of Y from **O**.

In a SEM with graph G, the *Markov blanket of X_i in G* (**MB**(X_i,G)) is the set of vertices that consists of the parents of X_i in G, the children of X_i in G, and the parents of the children of X_i in G. For example in SEM K, the Markov Blanket of R is A, and the Markov Blanket of A is $\{B, R\}$. If G contains a set of vertices **X** and represents the population density, then X_i is independent of all of the vertices not in its Markov Blanket

conditional on its Markov blanket, i.e. $I(X_i,\mathbf{X}\backslash\mathbf{MB}(X_i,G)|\mathbf{MB}(X_i,G))$. This implies that if X_i is regressed on a set of variables $\mathbf{O}$ that properly includes $\mathbf{MB}(X_i,G)$, then the regression coefficients of the variables outside of the Markov Blanket will be exactly zero. Since the results of regressing X_i on the entire set of variables $\mathbf{O}$ cannot be improved on with respect to MSPE, and regressing X_i on the Markov Blanket produces the same result as regressing X_i on the entire set of variables $\mathbf{O}$, regressing X_i on the Markov Blanket cannot be improved on (if G represents the population joint density).

For example, in K the Markov Blanket of R is A, and $I_K(R,B|A)$. Because the population covariance matrix is used in calculating the partial regression coefficients, the coefficient for the value of B when Y is regressed on A and B is exactly zero. Hence, given the population covariance matrix there is no difference between regressing Y on A (the Markov Blanket of R) and regressing Y on A and B (the set of all other variables).

If SEM G does not represent the population density, then regressing on a set of variables that properly contains the Markov Blanket of G may produce a better predictor of Y than regressing just on the Markov Blanket. Also, the fact that regressing on the Markov Blanket when given the population density produces a predictor with the smallest MSPE does not imply that regressing on the Markov Blanket when given a sample from the population distribution produces a predictor with the smallest MSPE.

3.2. Predicting Values From Samples

Suppose that instead of being given a population covariance matrix, a sample of size n from the population density $f(\mathbf{O},Y)$. In this case $f(\mathbf{O},Y)$ is not known, and so a predictor $\hat{Y}(\mathbf{O})$ cannot constructed from the population density. The goal is to find a function that maps samples of size n to a predictor $\hat{Y}_n(\mathbf{O})$ that has a low MSPE.

Problem 2: Sample predictive modeling

Input: A sample S of size n containing i.i.d. $\{\mathbf{o}, y\}$ samples drawn from $f(\mathbf{O}, Y)$ (i.e. training data), and a target variable Y to be predicted.

Output: $\hat{Y}_n(\mathbf{O})$, a predictor of Y that has a small MSPE.

3.2.1 Example

The following illustrates one approach (among several which have various advantages and disadvantages) for producing $\hat{R}_n(\{A,B\})$, and will be described in more detail in subsequent sections. The basic idea is to predict R by finding a good estimate of $E_f(R|\{A,B\})$, which can be solved by finding good estimates of the population partial regression coefficients. If the estimated values of the partial regression coefficients are not far from the population partial regression coefficients, then the MSPE will be guaranteed to be close to that of $E_f(R|\{A,B\})$, which is optimal.

Step 1: Search for a statistical model with a good score. This involves two problems: searching the space of models, which might be very large, and scoring models. The problem of searching the space of statistical models will not be discussed here. What constitutes a good score for Problem 2, and the reason for searching for models with good scores is described in section 3.2.2. In this example, suppose that no SEM receives a better score than L.

Step 2: Estimate the values of the free parameters of L.

Step 3: Use the estimates of the free parameters of L to calculate, $\hat{r}_{R,A.\{A,B\}}$ and $\hat{r}_{R,B.\{A,B\}}$, estimates of $r_{R,A.\{A,B\}}$ and $r_{B,A.\{A,B\}}$, respectively, by using the estimates of the free parameters to calculate the estimated covariance matrix, and using the estimated covariance matrix to calculate an estimate for the partial regression coefficients using Equation 3-1, as shown in

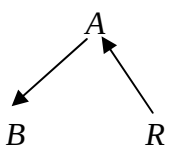
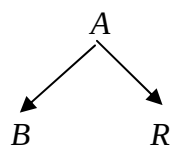
Figure 6. Hence

$$\hat{R}_n(\{A=a,B=b\}) = \hat{E}(R \mid \{A=a,B=b\}) =$$

$$E_{L(\hat{\boldsymbol{\theta}}')}(R) + \mathrm{cov}_{L(\hat{\boldsymbol{\theta}}')}(R,\{A,B\})\mathrm{var}_{L(\hat{\boldsymbol{\theta}}')}(\{A,B\})^{-1}\begin{bmatrix} a - E_{L(\hat{\boldsymbol{\theta}}')}(A) \\ b - E_{L(\hat{\boldsymbol{\theta}}')}(B) \end{bmatrix} =$$

$$E_{L(\hat{\boldsymbol{\theta}}')}(R) + \begin{bmatrix} \hat{r}_{R,A.\{A,B\}} & \hat{r}_{R,B.\{A,B\}} \end{bmatrix}\begin{bmatrix} a - E_{L(\hat{\boldsymbol{\theta}}')}(A) \\ b - E_{L(\hat{\boldsymbol{\theta}}')}(B) \end{bmatrix}$$

where $\hat{\theta}'$ is the estimate of the values of the free parameters of θ', $\mathrm{var}_{L(\hat{\theta}')}$ is the entailed covariance matrix of , and $E_{L(\hat{\theta}')}$ is the entailed expected value.

Estimated *L*	**Estimated *K***	**Estimated Covariances (*K* and *L*)**
		
Equations $R = \varepsilon'_R$ $B = \hat{b}_{B,A} \cdot A + \varepsilon'_B$ $A = \hat{b}_{A,R} \cdot R + \varepsilon'_A$	**Equations** $A = \varepsilon_A$ $B = \hat{b}_{B,A} \cdot A + \varepsilon_B$ $R = \hat{b}_{R,A} \cdot A + \varepsilon_R$	
Parameter Values $\hat{\sigma}^2(\varepsilon_A) = .1932$ $\hat{\sigma}^2(\varepsilon_B) = .4302$ $\hat{\sigma}^2(\varepsilon_R) = 1.021$ $\hat{b}_{B,A} = .3398 \quad \hat{b}_{A,R} = .7849$	**Parameter Values** $\hat{\sigma}^2(\varepsilon_A) = .8221$ $\hat{\sigma}^2(\varepsilon_B) = .4302$ $\hat{\sigma}^2(\varepsilon_R) = .2400$ $\hat{b}_{B,A} = .3398 \quad \hat{b}_{R,A} = .9746$	

	A	*B*	*R*
A	.822		
B	.279	.525	
R	.802	.272	1.02

$\hat{r}_{R,A.\{A,B\}} = 0.975$
$\hat{r}_{R,B.\{A,B\}} = 0.0$

Figure 6: Estimated Parameters and Covariance Matrix for *K* and *L* (Sample 50)

One question that naturally arises is why the procedure described above first searches for a statistical model with a low score, and then estimates the values of the free parameters of the model. Why not simply use a model that contains all of the probability densities? For example, one estimator of *Y* simply uses Equation 3-1 but replaces the population expected values with the sample expected values, and the population covariance matrix with the sample covariance matrix (where the sample covariance matrix of a vector **X** is $\mathrm{cov}_S(\mathbf{X}) = \mathbf{X} \cdot \mathbf{X}^t/n\text{-}1$, *n* is the sample size, and if there are *m* variables in **X**, **X** is an $n \times m$ matrix of sample points).

Equation 3-2 $\hat{Y}(\mathbf{O} = \mathbf{o}) = \hat{E}(Y \mid \mathbf{O} = \mathbf{o}) = E_S(Y) + \mathrm{cov}_S(Y, \mathbf{O})\,\mathrm{var}_S(\mathbf{O})^{-1} \cdot (\mathbf{o} - E_S(\mathbf{O}))$

Estimating the sample covariance matrix is equivalent to estimating the covariance matrix for a statistical model that contains the entire family of Gaussian densities over **O**; the sample covariance matrix is the maximum likelihood estimator of the covariance matrix subject to no constraints. Equation 3-2 simply replaces the

population regression coefficients with the sample regression coefficients, which are a function of the sample covariance coefficients. This minimizes the sum of squared residuals on the sample S, but it does not necessarily minimize the MSPE (that is the mean squared out of sample error), for reasons related to overfitting, as explained below.

Using the sample covariance matrix and regressing *R* on all the other variables is equivalent to using a SEM with a *complete graph* (i.e. a graph in which every pair of vertices is adjacent) to predict *R*. (For example, SEM *M* of

Figure 7 has a complete graph.) This is because the sample covariance matrix is the entailed covariance matrix for $M(\hat{\alpha})$, where $\hat{\alpha}$ is the maximum likelihood estimate of the free parameters of *M*.

In contrast, SEMs *K* and *L* both have fewer free parameters than *M*, and impose the (true population) constraint that cov(*R*,*B*|*A*) = 0 on the estimated covariance matrix. On the same sample of size 50, the estimated covariance matrix for both SEMs *K* and *L* in

Figure 6 is the same as the sample covariance matrix, except that $\text{cov}_S(B,R) = .2093$. Using Equation 3-2, the estimated partial regression coefficients from the sample covariance matrix are $\hat{r}_{R,A.\{A,B\}} = 1.0243$ and $\hat{r}_{R,B.\{A,B\}} = -.1461$. The MSPE (estimated from a sample of 5000) is slightly higher for the partial regression coefficients estimated from the sample covariance matrix than for the partial regression coefficients estimated from *K* or *L* (.399 versus .395).

The explanation for the superiority, at some small sample sizes, of using *K* or *L* to predict *R* over using *M* to predict *R* is the following. If the mean square error (MSE) of the difference between the estimate $\hat{r}_{Y.\mathbf{O}}$ and the population $r_{Y.\mathbf{O}}$ (taken with respect to the sampling density of $P(\mathbf{O},Y)$) is small, then the MSPE for *Y* also has to be small. The mean square error of an estimator is the sum of the bias of the estimator and the variance of the estimator. Consider predicting *R* from *M*. The Markov Blanket of *R* in *M* is *A* and *B*. The

population regression coefficient for the value of *B* is zero (because it is not in the Markov blanket of *R* in the true SEM *K*), but the sample regression coefficient for the value of *B* is typically non-zero. In contrast, in *K* or *L* the Markov Blanket of *R* in *K* or *L* is *A*. Using *K* or *L* to predict *R* is tantamount to leaving *B* out of the regression (which is equivalent to setting the coefficient of *B* to zero).[14] Leaving *B* out of the regression also affects the estimate of the partial regression coefficient of *A*. Regressing *R* only on *A* will not increase the bias of the estimate of the partial regression coefficient, but it will decrease the variance of the estimator, leading to a smaller mean square error. In the case of Gaussian variables, the difference is usually small except at very small sample sizes or very large numbers of variables. For example, if there are more variables than sample points, there is no unique maximum likelihood estimate of the covariance matrix. For other parametric families the differences can be quite large even for large sample sizes. Even for cases where leaving variables out does bias the estimator (e.g. in cases where the population regression coefficient is close but not equal to zero) the corresponding decrease in the variance of the estimator may lead to a decrease in mean squared error at a given sample size.

In looking for an estimator that minimizes the MSE of the partial regression coefficients, one strategy that can produce estimators with a small MSE is to break Problem 2 into the several subproblems of steps 1 through 3 - finding a "good" statistical model, estimating the parameters of the model, and using the estimated model to construct an estimator of *Y*. Each of these subproblems will be discussed in the subsequent subsections.

3.2.2 Searching and Scoring Models

A statistical model is a set of probability densities. What constitutes a "good" statistical model depends upon what purpose the model is to be used for. Ideally, then a "good" model for Problem 2 is one, which when its free parameters are estimated,

---

[14] In this simple example, using the estimated covariance matrix from SEMs *K* or *L* is equivalent to using the sample covariance matrix, but regressing *R* only on *A*. However, in general, using the estimated covariance matrix from a SEM does not correspond to using the sample covariance matrix and regressing on some subset of the variables, because a SEM can impose constraints on the covariance matrix that affect more than which variables appear in the Markov Blanket.

produces a predictor $\hat{Y}_n(\mathbf{O})$ with a small MSPE. There are a variety of scores that are used to guess which models have this property. This section briefly discusses the Aikake Information Criterion (AIC) and cross-validation scores, in order to contrast them later with different scores that are more appropriate for causal models.

Problem 2a: Model Search
Input: A sample of n $\{\mathbf{o}, y\}$ samples drawn randomly and independently according to P($\mathbf{O}$, Y).
Output: A statistical model that minimizes a score that is a function of the data and the model.

(which is likely to lead to overfitting). In the case of a multi-variate Gaussian SEM M, for a given sample of size n

$$\text{AIC}(M\text{ , sample}) = -2 \cdot L(\text{var}_{M(\hat{\boldsymbol{\theta}})}, \text{sample}) + 2 \cdot \text{df}_M,$$

where

- $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate[15] of the parameters for model M from the sample; and
- $L(\Sigma_{M(\hat{\boldsymbol{\theta}})}, \text{sample})$ is the likelihood of $\text{var}_{M(\hat{\boldsymbol{\theta}})}$; and
- df_M is the degrees of freedom (dimensionality) of the SEM M.

Choosing a model with a minimum AIC score has several desirable properties, including the property that it minimizes the Kullback-Leibler[16] distance between a predicted density and the true density. It is commonly used when the goal is find a model that is good for predictive purposes, and does not assume that the true density is in any of the competing statistical models. Even if the true density is in one of the competing statistical models, it is not always the case that the statistical model with the lowest AIC score is a statistical model that contains the true density even at large sample sizes.

As discussed in section 5.2.2, SEMs K and L always have the same AIC score regardless of the data, because they represent the same set of (Gaussian) densities and

[15] Maximizing the likelihood is equivalent to minimizing certain risk functions, the more usual approach in machine learning problems.

[16] If the true density is f, and the predicted density is g, then the KL distance is $\int f \cdot \log(f/g)$

have the same number of parameters. Hence, AIC does not help distinguish between them. This is not a problem as long as the goal is prediction, because using SEM *K* instead of SEM *L* to predict *R* leads to the same predictor.

In the case of *L*, $\mathrm{cov}_{L(\hat{\boldsymbol{\theta}}')}(B, R \mid A) = 0$, and hence $\mathrm{var}_{L(\hat{\boldsymbol{\theta}}')}$ is not in general equal to the sample covariance matrix. As noted in section 3.1, in any density represented by a SEM, regressing *R* on any superset of the variables in the Markov Blanket of *R* in the graph of the SEM guarantees that the variables outside of the Markov Blanket have estimated partial regression coefficients equal to 0. In the case of *L*, the Markov Blanket for *R* is *A*, and hence, and $\hat{r}_{R,B.\{A,B\}} = 0$. Hence whether *B* is a child of *A* or not makes no difference to the prediction of *R*.

In some cases, the SEM that is the best predictor of *Y* at small sample sizes is a SEM with unobserved variables. According to SEM *C*, there are no conditional independence relations among just the variables in $\mathbf{O} = \{X_1, X_2, X_3, X_4\}$, so the only SEMs without hidden variables that represent the marginal density over **O** all are complete graphs. For example, *D*, where $\mathbf{O} = \{X_1, X_2, X_3, X_4\}$, is a complete DAG.

However, there are constraints on the covariance matrix among the variables in **O** entailed by *C* ("vanishing tetrad constraints" (Harman, 1976)), which are not entailed by any SEM *D* or any SEM with a complete graph. In *C*, applying the trek rule, $\mathrm{cov}_C(X_1,X_2) \cdot \mathrm{cov}_C(X_3,X_4) = \mathrm{cov}_C(X_1,X_3) \cdot \mathrm{cov}_C(X_2,X_4) = \mathrm{cov}_C(X_1,X_4) \cdot \mathrm{cov}_C(X_2,X_3)$. So the set of marginal densities over $\{X_1, X_2, X_3, X_4\}$ represented by *C* has lower dimension than the set of probability densities over the same variables represented by any SEM without an unmeasured variable, such as *D*.[17] Because it is simpler but still represents the population density, at small sample sizes *C* tends to have a lower AIC score and a lower MSPE than any of the alternative SEMs with complete graphs over **O**. However, alternative SEMs that represent the population density will still provide unbiased estimates of conditional expected values, even if they have a higher variance and a higher MSE.

---

[17] Using AIC to select a SEM that can be used to construct an estimator of *Y* with a low MSPE is not guaranteed to produce the best results for a number of reasons, including the fact that AIC measures the simplicity of the entire model, whereas the simplicity of *Y* conditional on the predictors is more directly relevant.

A SEM can also be selected by its cross-validation score. In a k-fold cross-validation, the data is divided into a k subsets. For each subset S, the other k-1 subsets are used as a training set to generate a SEM with fixed parameters. Then the SEM with fixed parameters is used to predict the value of the target variable in S, and the sample MSPE on S is calculated. The SEM with the lowest sample MSPE is selected.

However, both AIC and k-fold cross-validation have theoretical problems when employed as a score for causal models, although in some circumstances they may perform well. See section 7.1.1.

3.2.3 Model Estimation

Because the estimate of R is a continuous function of the partial regression coefficients, as long as the partial regression coefficient estimates have small mean square error, so will the estimate of R.

Since L contains no unobserved common causes, a maximum likelihood estimate of its linear coefficients b_{ij} is equal to the partial regression coefficient of X_j for X_i when X_i is regressed on its parents in L, using the sample covariance matrix as input. This can be computed analytically. Otherwise, an iterative algorithm that takes the sample covariance matrix as input is used to find maximum likelihood estimates of the linear coefficients. The result is $L(\hat{\boldsymbol{\theta}}')$ where $\hat{\boldsymbol{\theta}}'$ is the maximum likelihood estimate of the free parameters of L. In some SEMs with hidden common causes, there are no consistent estimators of the free parameters of the model. The question of whether the free parameters of a SEM with hidden common causes are identifiable is quite difficult to solve; recently, algebraic methods have solved the problem for some families of densities, including Gaussian (Geiger & Meek, 1999). However, these algorithms are computationally quite intensive, and are only practical for a few variables. In those cases where the free parameters of a SEM with hidden common causes are not identifiable, it is always possible to use SEMs without hidden common causes to construct $\hat{R}_n(\{A, B\})$.

Problem 2b: Model Estimation

Input: A sample of n $\{\mathbf{o}, y\}$ examples drawn randomly and independently according to the sampling distribution of $f(\mathbf{O}, Y)$, and a statistical model.

Output: A consistent estimate of the values of the free parameters of the model.

4. Predicting Manipulated Values of *Y*

There are causal estimation problems that are analogous to the various statistical estimation problems described in the previous section.

4.1. Causal Inference Given the Population Density

Problem 3: Probabilistic causal predictive modeling

Input: $f(\mathbf{O},Y)$, a manipulation $f'(X_i=x_i)$ of X_i, the post-manipulation values **o'** of a subset **O'** of the variables **O**, and a target variable Y whose post-manipulation value is to be predicted.

Output: $\hat{Y}$ (**O'=o'**, $f'(X_i=x_i)$), a predictor of the post-manipulation value of Y that has a small MSPE.

In both Problems 1 and 3, the unmanipulated density is given, and Y is to be predicted. The major difference is that in Problem 3 the units for which Y is to be predicted are drawn from the manipulated density (as opposed to the given unmanipulated density). It is assumed that some (possibly empty) set of post-manipulated values for a subset of **O'** of **O** is known. The post-manipulation values of **O'** could be known either by performing a post-manipulation measurement of the values, or through background knowledge, or through inference from the given unmanipulated probability density $P(\mathbf{O},Y)$. For example, in some cases even if the pre-manipulation values of the barometer reading and atmospheric pressure are measured, the post-manipulation values of barometer reading and atmospheric pressure are not. In other cases, the atmospheric pressure and barometer reading could be measured after the manipulation. Alternatively, if it is known that the barometer reading does not affect the atmospheric pressure, then the atmospheric pressure after manipulating the barometer reading is known because it is the same as the atmospheric pressure prior to manipulating the barometer reading (assuming the system is in a steady state).[18]

[18] Here the sense in which the barometer reading does not affect the atmospheric pressure should be understood as implying that no individual barometer reading affects any individual atmospheric pressure, not simply that the density of barometer readings does not affect the density of atmospheric pressures.

The difference between the way these problems are solved will be illustrated mainly by considering predicting R under various combinations of manipulating and observing A and B. The basic strategies have one major difference from strategies for predicting an unmanipulated value of R. Given the population density f, when predicting R from observed A and B, $E_f(R|A,B)$ is used to make the prediction. In predicting R when $f(A)$ is manipulated to $f'(A)$ and $f(B)$ is manipulated to $f'(B)$ (but the post-manipulation values of A and B are not observed), the expected value of R in the manipulated population, $E_f(R||f'(A),f'(B))$, is used to make the prediction. (Note that if A and B had been manipulated to some new density, and their post-manipulation values observed, $E(R|A,B||f'(A),f'(B))$ would be used to make the prediction.

4.1.1 An Example

Given the population density, the following steps can be used to calculate $E_f(R||f'(A),f'(B))$. The steps will be elaborated on and justified in more detail in the following sections.

Step 1: Find the causal models "compatible" with background knowledge and the population density. (There are several different senses of "compatible", discussed in more detail below.) Suppose that the true but unknown SEM is K. Background knowledge that A occurs before R, and that the set of variables is causally sufficient is given. Then there are two causal models "compatible" with the population density and background knowledge are K and M.

Step 2: Find an expression for $E(R|||f'(A),f'(B))$ as a function of population conditional probabilities and the manipulation that is the same for both candidate SEMs K and M. As shown in section 6,

$$E_K(R \parallel f'(A), f'(B)) = E_M(R \parallel f'(A), f'(B)) = \int_R R \cdot \int_A f(R \mid A) f'(A) dR dA$$

Since the two expressions are the same, proceed to step 3. If no such expression for $E(R|||f'(A),f'(B))$ which is the same for all candidate models can be found, return "Don't know".

Step 3: Arbitrarily choose one of the candidate models, e.g. SEM M, calculate $f_M(R|A)$ and substitute it into the equation in step 2. Return the predicted value.

5. Finding Compatible Causal Models

There are two major differences between model search in the case of predicting the unmanipulated value of *Y*, and model search in the case of predicting the post-manipulation value of *Y*. The first difference is that model search in the case of predicting the unmanipulated value of *Y* is only useful when the input was a sample from a population density, not the population density itself. In contrast, in the causal case, even if the population density is given, model search is still necessary. The second difference is that in the case of predicting the unmanipulated value of *Y*, the model search is over a space of statistical models, i.e. sets of probability densities for unmanipulated populations. In the case of predicting manipulated values of *Y*, the model search is over a space of causal models, i.e. sets of probability densities for both unmanipulated and post-manipulation populations.

Problem 3a: Causal Model Search

Input: The density $f(\mathbf{O},Y)$.

Output: The set of causal models with free parameters that are compatible with $f(\mathbf{O},Y)$.

The conditional expected value is a function of the population joint probability density.

$$E_f(R \mid A,B) = \int R \frac{f(R,A,B)}{f(A,B)}$$

The post-manipulation expected value is not a function of the population joint probability density alone: two populations can have the same joint density, and different post-manipulation expected values. Two SEMs (with free parameters) are *Markov equivalent* if they entail the same set of conditional independence relations (or equivalently, for SEMs with acyclic graphs, the acyclic graphs have the same set of d-separation relations.) Two SEMs (with free parameters) are *distribution equivalent* if and only they represent the same set of distributions. (Distribution equivalence implicitly refers to a parametric family, e.g. linear models, or in this example, Gaussian models.)

For example, graph *K* has the same d-separation relations as graphs *L* and *M*, and hence entails the same conditional independence relations as graphs *L* and *M*. In addition,

SEM K (with free parameters) represents the same set of Gaussian densities as SEMs L and M (with free parameters) and hence K, L, and M are *distribution equivalent*. (If the equations are linear, but non-Gaussian error terms are allowed, the corresponding SEMs are Markov equivalent, but they are not distribution equivalent.)

Given the right choice of parameter values, as in Figure 2 and Figure 4, SEMs K and L can represent the same covariance matrix. Suppose that $K(\theta)$ and $L(\theta')$ entail the same covariance matrix, and that $E(f'(A)) = a \neq 0$. By the linearity of expectation, $E_{K(\theta)}(R||f'(A),f'(B)) = b_{R,A} \cdot a = \text{cov}(R,A) \cdot a = E_{K(\theta)}(R|A=a) \neq E_{K(\theta)}(R) = 0$. (Recall that $b_{R,A}$ is the partial regression coefficient for A when R is regressed on A, and that when A is regressed on a single variable with variance 1 that the partial regression coefficient is equal to cov(R,A).) In K, every trek between A and R (which determines the covariance) is also a directed path from A to R (which determines the total effect), so the effect of manipulating is the same as that of conditioning.

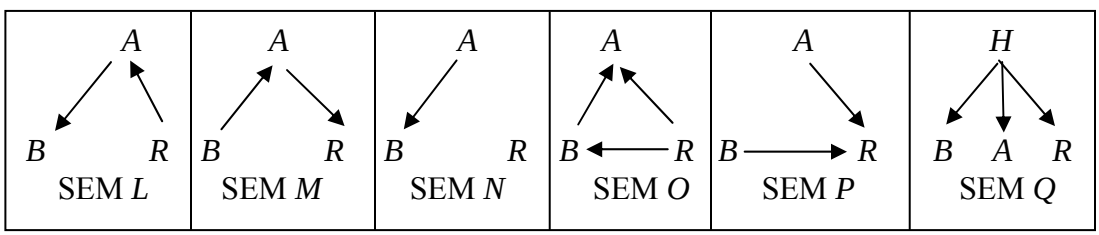


Figure 7: Alternative SEMs

In contrast, $E_{L(\theta')}(R||f'(A),f'(B)) = E_{L(\theta')}(R) = 0 \neq E_{L(\theta')}(R|A=a) = b_{A,R} \cdot a = \text{cov}(R,A) \cdot a$. In L there is a trek between A and R ($A \leftarrow R$) that is not a directed path from A to R, and hence the conditional expected value (determined by the trek sum) is not equal to the manipulated expected value (determined by the path sum).

This shows that in general, the manipulated expected value is not a function of the population density alone - it also depends upon features of the causal relationships between the variables, as represented in the causal graph.

One feature of the causal graph that affects the manipulated expected value is the causal order, i.e. whether A is a cause of R, or an effect of R, or neither. A second feature of the causal graph that affects the manipulated expected value is confounding, or common causes. For example, $E_{K(\boldsymbol{\theta})}(R||B=m) = 0 \neq \text{cov}_{K(\boldsymbol{\theta})}(R,B) = b_{R,A} \cdot b_{B,A}$, because there is a trek between R and B ($B \leftarrow A \rightarrow R$) but no directed path from B to A.

5.1. Assumptions Relating Causal Models to Population Densities

As indicated by SEMs *K*, *L*, and *M*, there can be more than one SEM that is compatible with a given population density. However, if it is possible to use the population density to eliminate any causal models as false, there must be assumptions that relate the causal models to (unmanipulated) population densities. The following two sections describe two assumptions that are commonly, but often implicitly, made that relate causal models to population densities.

5.1.1 The Causal Markov Assumption

The following assumption is used to relate causal relations to probability densities.

Weak Causal Markov Assumption: For a causally sufficient set of variables **V** in a population *N*, if no variable in **X** causes any variable in **Y**, and no variable in **Y** causes any variable in **X**, then **X** and **Y** are independent (i.e. in the Gaussian case, members of **X** and members of **Y** are pairwise uncorrelated.) (Spirtes et al., 2001)

The Weak Causal Markov Assumption does not hold for non-causally sufficient sets of variables. For example, if *K* is true, then {*B*, *R*} is not causally sufficient. Despite the fact that *B* does not cause *R* and *R* does not cause *B*, *R* and *B* are dependent (because of the common cause *A* that is not in the set of variables considered.)

The Weak Causal Markov Assumption has the consequence that the error terms for causally sufficient sets of variables are independent. Simon's famous analysis of "spurious correlation" (Simon, 1985) is precisely an application of the Weak Causal Markov Assumption to explain correlated errors as the result of unobserved common causes. The examples that Bollen gives of why an error 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 error term and other causes of *X*, and hence an application of the Weak Causal Markov Assumption (Bollen, 1989). (For a discussion of the Causal Markov Assumption, and conditions under which it should not be assumed, see Spirtes et al., 2001).

For deterministic causal models such as SEMs, the Weak Causal Markov Assumption also entails another version of the Causal Markov Assumption, i.e. that for causally sufficient sets of variables, all variables are independent of the their non-effects (non-

descendants in the causal graph) conditional on their direct causes (parents in the causal graph) (Spirtes et al., 2001).[19]

Although the Weak Causal Markov Assumption only applies to causally sufficient sets of variables, some of the causal discovery algorithms described later that use the Causal Markov Assumption merely assume that causally sufficient sets of variables *exist*, not that the causally sufficient sets of variables are all *observed*, i.e. there can be unobserved common causes.

The Weak Causal Markov Assumption is an oversimplification because it basically assumes that all associations between variables are due to causal relations. There are several other ways that associations can be produced.

First, conditioning on a common descendant can produce a conditional dependency. For example, if sex and intelligence are unassociated in the population, but only the most intelligent women attend graduate school, while men with a wider range of intelligence attend graduate school, then sex and intelligence will be associated in a sample consisting of graduate students (i.e. sex and intelligence cause graduate school attendance, which has been conditioned on in the sample.) See (Spirtes, Meek, & Richardson, 1995) for a discussion of selection bias. Second, logical relationships between variables can also produce non-causal correlations (e.g. if *GDP_yearly* is defined to be the sum of *GDP_January*, *GDP_Februrary*, etc., *GDP_yearly* will be associated with these variables, but not caused by them.) For a discussion of logical relations between variables, see (Spirtes & Scheines, 2004). Third, it does not have any way of dealing with instantaneous symmetric interactions (like classical theories of gravity).

The Causal Markov Assumption entails that some causal models are incompatible with some population densities, and hence can be used to eliminate some causal models as false. For example, in *K*, every linear parameterization entails that *A* and *R* are dependent. So if *K* is the true SEM, any SEM (such as SEM *N* in Figure 7) with no trek between *A* and *R* entails that *A* and *R* are independent and so can be eliminated as false.

Unfortunately, given a population density, the Causal Markov Assumption does not eliminate enough alternative causal models to be able to draw any reliable conclusions

[19] For non-deterministic causal models, the alternative Causal Markov Assumption is usually made directly. (For different alternative versions of Markov relations, see Lauritzen et al., 1990).

about the effects of manipulations. In particular, regardless of what the population density is, every possible causal ordering of the variables is compatible with the some DAG G for which each vertex is independent of its non-descendants in G conditional on its parents in G. (For example, a complete graph with any ordering of the variables entails no conditional independence relations and so trivially satisfies the Causal Markov Assumption.) Some further assumptions relating causal models to population probability densities will now be considered.

5.1.2 The Causal Faithfulness Assumption

Consider SEM O in

Figure 7. $I_K(B, R|A)$, whereas it is not the case that $I_O(B,R|A)$. However, just because O does not entail $I_{O(\theta)}(B,R|A)$ for all sets of parameter values β, that does not imply that there are *no* β for which $I_{O(\beta)}(B,R|A)$. For example, if the variances of R, A, and B are all 1, for any β for which $\text{cov}_{O(\beta)}(A,B) \cdot \text{cov}_{O(\beta)}(A,R) = \text{cov}_{O(\beta)}(B,R)$, it follows that $\text{cov}_{O(\theta)}(B,R|A) = 0$. By the trek rule, this occurs when $(b_{B,R} \cdot b_{A,R} + b_{A,B}) \cdot (b_{B,R} \cdot b_{A,B} + b_{A,R}) = b_{R,B}$. So if $I_f(B,R|A)$ is true in the population, there are at least two kinds of explanation: any set of parameter values for SEMs K, L, or M on the one hand, or any parameterization of SEM O for which $(b_{B,R} \cdot b_{A,R} + b_{A,B}) \cdot (b_{B,R} \cdot b_{A,B} + b_{A,R}) = b_{R,B}$. There are several arguments why, although O with the special parameter values is a possible explanation, in the absence of evidence to the contrary, K, L, or M should be the preferred explanations.

K, L, and M explain the independence of B and R conditional on A structurally, as a consequence of no direct causal connection between the variables. In contrast O explains the independence as a consequence of a large direct effect of B on R cancelled *exactly* by the product of large direct and indirect effects of B and R on A. The latter explanation is contrary to the practice of many sciences, which typically assumes that, unless there is evidence to the contrary, an improbable and unstable cancellation of parameters (as in O) does not hide real causal influences (such as the causal influence of B on R in O). 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 (Glymour, 1980).

Second, this cancellation is improbable (in the Bayesian sense that if a zero conditional covariance is not entailed, the measure of the set of free parameter values for any DAG that lead to such cancellations is zero for any "smooth" prior probability density[20] e.g. Normal, exponential, etc., over the free parameters).

Finally, *K*, *L*, and *M* are simpler than *O*. *K*, *L*, and *M* have fewer free parameters than *O*. Because *O* imposes a proper subset of the conditional independence constraints imposed by *K*, *L*, and *M*, *O* represents a proper superset of the densities that can be represented by *K*, *L*, and *M*. There is a well defined sense in which the set of densities represented by *O* is of higher dimension than the set of densities represented by *K*, *L*, and *M* (Geiger, Heckerman, King, & Meek, 2002). So *O* is more complex than *K*, *L*, and *M* in a precisely defined way. This is in accord with all of the usual scores for statistical models (Bayes Information Criterion, X^2, Aikake Information Criterion, etc.), which, given a density represented by *K*, prefer SEMs *K*, *L*, and *M* to SEM *O* because of their greater simplicity.

The assumption that a causal influence is not hidden by coincidental cancellations can be expressed for SEMs in the following way. A density *f* is *faithful* to the graph *G* of a SEM if and only if every conditional independence relation true in *f* is entailed by *G*.

Causal Faithfulness Assumption: For a causally sufficient set of variables, the population density is faithful to the causal graph. (Spirtes et al., 2001)

The Causal Faithfulness Assumption requires preferring *K*, *L*, and *M* to *O*, because parameter values $\boldsymbol{\beta}$ for which $I_{O(\boldsymbol{\beta})}(B,R|A)$ would violate the Causal Faithfulness Assumption. The Causal Faithfulness Assumption limits the SEMs considered to those SEMs in which population conditional independence constraints are entailed by causal structure, rather than by particular values of the parameters. There can be sometimes be good reasons to believe that the more complicated model is true, and the reason that a conditional independence relation holds is not due to the structure of the graph, e.g. when there are deterministic relationships among the substantive variables, or equality

[20] A smooth measure is absolutely continuous with Lebesgue measure.

constraints upon free parameters. In those cases, the Causal Faithfulness Assumption should not be made (Spirtes et al., 2001).

There are several other versions of the assumption that are considerably weaker than the one stated here (and more intuitively justifiable) but still permit reliable causal inference, at the cost of requiring more complicated algorithms with more complex and somewhat less informative output (Zhang & Spirtes, 2008; Ramsey, Spirtes, & Zhang, 2006)Ramsey et al., 2006, #177].

5.2. The Output of A Search for Causal Models

The following sections describe several different possible alternatives that can be output by a reliable search algorithm.

5.2.1 Markov Equivalence Classes

SEMs *K*, *L*, and *M* are Markov equivalent. If *K* is true, any SEM with no trek between *A* and *R* can be eliminated from consideration by the Causal Markov Assumption (e.g. *N*). *P* also violates the Causal Markov Assumption. *O* is incompatible with the population conditional independencies by the Causal Faithfulness Assumption. However, neither of these assumptions implies *L* or *M* is incompatible with the population conditional independencies.

Since *K*, *L*, and *M* entail the same set of conditional independence relations, it is not possible to eliminate *L* or *M* as incompatible with the population conditional independence relations without either adding more assumptions or background knowledge, or using features of the probability density that are not conditional independence relations. In the case of Gaussian error terms (and for multinomial Bayesian networks) there are no other features of the density that distinguish *K* from *L* or *M*.

A collider $X \rightarrow Y \leftarrow Z$ is *unshielded* if *X* and *Y* are not adjacent. Two DAGs are Markov equivalent if and only if they have the same adjacencies, and the same unshielded colliders (Verma & Pearl, 1990). Any subset *S* of a Markov equivalence class of DAGs can be represented by a *pattern* (also known as a *pdag* or *essential graph*). A *pattern* has the same adjacencies as each member of the Markov equivalence class, and a directed edge $X \rightarrow Y$ if and only if each member of *S* has a directed edge $X \rightarrow Y$, and an

undirected edge $X - Y$ if and only some member of S has a directed edge $X \rightarrow Y$, and some other member of S has a directed edge $X \leftarrow Y$. For example, the Markov equivalence class of K is represented by $B - A - R$. The Markov equivalence class of P is represented by $B \rightarrow R \leftarrow A$ (i.e. it has only one DAG in the equivalence class.) Other DAGs can have mixtures of directed and undirected edges (e.g. $X - B \rightarrow A \leftarrow R$ is also a pattern that represents two DAGs, $X \rightarrow B \rightarrow A \leftarrow R$ and $X \leftarrow B \rightarrow A \leftarrow R$).

5.2.2 Distribution equivalence

K and L are *distribution equivalent* if and only if for any assignment of parameter values θ to K there exists an assignment of parameter values θ' to L that represents the same density, and vice versa. If all of the error terms are Gaussian, then K and L are distribution equivalent as well as Markov equivalent. In such cases, the best that a reliable search algorithm can do is to return the entire Markov equivalence class, regardless of what features of the marginal density that it uses.

In contrast, if at most one error term is non-Gaussian, SEMs K and L are Markov equivalent, but they are not distribution equivalent.

When Markov equivalence fails to entail distribution equivalence, then using conditional independence relations alone for causal inference is still correct, but it is not as informative as theoretically possible. For example, assuming causal sufficiency and non-Gaussian errors (Shimizu, Hoyer, Hyvarinen, & Kerminen, 2006), conditional independence tests can at best reliably determine the correct Markov equivalence class, while using other features of the sample density can be used to reliably determine a unique graph (Shimizu et al., 2006).

5.3. Constraint-Based Search

The number of DAGs grows super-exponentially with the number of vertices, so even for modest numbers of variables it is not possible to examine each DAG to determine whether it is compatible with the population density given the Causal Markov and Faithfulness Assumptions. The PC algorithm, given as input an oracle that returns answers about conditional independence in the population and optional background knowledge about orientations of edges, returns a pattern that represents a Markov equivalence class (or if there is background knowledge a subset of a Markov equivalence

class) on the basis of oracle queries. If the oracle always gives correct answers, and the Causal Markov and Causal Faithfulness Assumptions hold, then the output pattern contains the true SEM, even thought the algorithm does not check each DAG. In the worse case, it is exponential in the number of variables, but for sparse graphs it can run on hundreds of variables (Spirtes & Glymour, 1991; Spirtes et al., 1993; Meek, 1995).

In contrast to the case where a statistical model is selected in order to predict the value of a variable, it is not possible to use cross-validation in the same way to test whether a causal model selected by a search is a good predictor of the manipulated value of a variable. If the data is repeatedly divided into a training set used to construct a model, and the constructed models greatly differ in their predictions of the manipulated value of a variable, then that is evidence that the search is not working well. However, if the constructed models differ only slightly in their predictions of the manipulated value of a variable, it is still possible that they are all far off in their predictions, because the manipulated value of the variable is typically not known without experimentation.

However, it is often difficult to experimentally test the output of causal search algorithms precisely because in many domains experimental confirmation is expensive and difficult to obtain. A number of causal inference methods have been extensively tested on simulations, but their experimental validation on empirical data is much more scattered (Sachs et al., 2005; Scheines, Leinhardt, Smith, & Cho, 2005; Spirtes et al., 1993). A causality repository, containing both simulated and real data sets that can be used for evaluating causal inference methods, together with links to some causal discovery methods is being developed at http://www.causality.inf.ethz.ch/repository.php.

6. Using Search Output to Calculate The Value of a Manipulated Variable

If the causal relations are assumed to be linear and at most one error term is Gaussian, the output of a LiNGAM search is a unique DAG with estimated linear coefficients (Shimizu et al., 2006). The value of any manipulation can then always be calculated by replacing the original structural equation with the new structural equation, and calculating the new density, as described in section 2.4. However, this method of calculating the new density has two major limitations in other cases. First, it does not always work if there are

unobserved common causes because only the marginal density over the observed variables is given. Second, it may be computationally infeasible if the output contains many DAGs compatible with the population density and background knowledge. For that reason, a different method that can be more easily extended to cases where there are unmeasured common causes, or the output is a pattern will be described here. This method is equivalent to Pearl's do-calculus (Pearl, 2000) and was described in Causation, Prediction, and Search (Spirtes et al., 1993).

6.1. Causal Inference from a Single SEM

For example, suppose that the only SEM compatible with the population density and background knowledge is K, and in order to predict R, the goal is to calculate $E_M(R||f'(A),f'(B))$ as a function of the unmanipulated population density, and the manipulation.

$$E_K(R \,||\, f'(A), f'(B)) = \int_R R \cdot f_K(R \,||\, f'(A), f'(B))dR$$

The problem with using this expression is that $f_K(R||f'(\mathrm{A}),f'(B))$ is a manipulated density, rather than an unmanipulated density, and $f_K(R||f'(\mathrm{A}),f'(B)) \neq f_K(R)$. However, since for any density $f(R) = \int_A f(R \mid A) f(A) dA$, and $f(R||f'(A),f'(B))$ is a density

$$\int_R R \cdot f(R \,||\, f'(A), f'(B))dR = \int_R R \cdot \int_A f(R \mid A \,||\, f'(A), f'(B)) f(A \,||\, f'(A), f'(B))dAdR$$

By definition, after manipulating the density of A to $f'(A)$, $f(A||f'(A),f'(B))$ is $f'(A)$, so

$$\int_R R \cdot \int_A f(R \mid A \,||\, f'(A), f'(B)) f(A \,||\, f'(A), f'(B))dRdA =$$

$$\int_R R \cdot \int_A f(R \mid A \,||\, f'(A), f'(B)) f'(A)dRdA$$

The latter expression is still not entirely in terms of the unmanipulated density and the given manipulation because it contains $f(R|A||f'(A),f'(B))$. However, the conditional density is *invariant under manipulation* $\{f(A), f'(B)\}$ if $f(R|A||f'(A),f'(B)) = f(R|A)$, that is the unmanipulated conditional density equals the manipulated conditional density. If $f(R|A||f'(A),f'(B))$ is invariant under manipulation, then $f(R|A)$ can be substituted into the expression for $f(R|A||f'(A),f'(B))$, and the expression will contain only the unmanipulated

(conditional) density and the manipulation. If $f(R|A)$ is invariant under manipulation of A and B, then $E(R||f'(A),f'(B))$ is equal to

Equation 6-1

$$\int_R R \cdot \int_A f(R \mid A \,\|\, f'(A), f'(B)) f(A \,\|\, f'(A), f'(B)) dRdA =$$
$$\int_R R \cdot \int_A f(R \mid A \,\|\, f'(A), f'(B)) f'(A) dRdA = \int_R R \cdot \int_A f(R \mid A) f'(A) dRdA$$

There is an easy graphical method to check whether $f(R|A)$ is invariant under manipulation $\{f'(A),f'(B)\}$. A manipulation in a causal model K can be represented by an augmented graph K'. If A is manipulated to have density $f'(A)$, this can be represented by introducing a new $Policy_A$ variable that has no edges coming into it, and one edge from $Policy_A$ to A. (The case where manipulations introduce new causes of a variable will not be considered here.) The values that this variable takes on can represent the value of the intervention - in this case a vector of values that represents the expected value and variance of ε_A, and the value of the linear coefficients of its parents in K, which in this case have been reset to zero after the manipulation. The manipulation of B can be represented similarly. Then define a new density f' over the extended set of variables including $Policy_A$ and $Policy_B$, so that by definition $f'_K(A, B, R|Policy_A = f'(A), Policy_B = f'(B)) = f_K(R||f'(A),f'(B))$; i.e. $f'_K(A, B, R|Policy_A = f'(A), Policy_B = f'(B))$ is the density that results from replacing the structural equations for A with $A = f'(A)$ and $B = f'(B)$. There is no need to define a joint density over the augmented set of variables; the only densities that are used are densities conditional on the policy variables. When $Policy_A$ and $Policy_B$ are set to the values in the unmanipulated SEM, then $f'_K(A,B,R|Policy_A, Policy_B) = f_K(A, B, R)$.

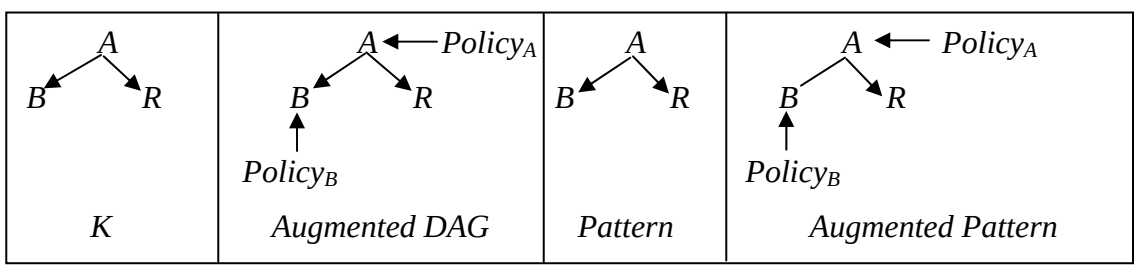


Figure 8: Graphs and Augmented Graphs

In the augmented graph, $Policy_A$ and $Policy_B$ are d-separated from R conditional on A. That implies $f'_K(A|R,Policy_A) = f'_K(A|R)$; in particular the manipulated and unmanipulated densities of $f'_K(A|R)$ are the same, so $f_K(A|R)$ is invariant under manipulation. In general if

the policy variables representing the densities are d-separated from **X** conditional on **Y**, then $f(\mathbf{X}|\mathbf{Y})$ is invariant under manipulation.

To summarize, the algorithm consists of searching for an expression for the desired quantity (in this case $f(R||f'(A),f'(B))$) in terms of arguments that are either functions of the manipulations $f'(A)$ and $f'(B)$, or of invariant manipulated conditional densities, (in this case $f(R|A||f'(A),f'(B))$). Invariance of $f(R|A||f'(A),f'(B))$ under manipulation is tested by determining if $Policy_A$ and $Policy_B$ are d-separated from R conditional on A in the augmented graph. If invariance does hold, then $f(R|A)$ can be substituted in for $f(R|A||f'(A),f'(B))$), turning the expression into one in which all of the arguments are either the given unmanipulated density ($f(R|A)$), or the manipulation ($f'(A)$). This algorithm is not particularly useful for the case where the output is a DAG with no hidden common causes, because it requires searching for the right expression, unlike the algorithm that simply replaces the structural equations for A and B. However, it does generalize to other cases described below.

6.2. Causal Inference from a Pattern

If the output of a search is a pattern (representing a subset of a Markov equivalence class) that contains more than one DAG, the value of a manipulated quantity may or may not be a function of the output and the population density, depending upon what manipulated quantity is being calculated, and what the true DAG is. For example, without any background knowledge about the causal graph (other than acyclicity), but assuming a linear SEM with Gaussian errors, in the large sample limit the output of the PC search will be $B - A - R$, if K is the true SEM. $E(R||f'(A),f'(B))$ is not a function of the population density and the pattern, because as already indicated in section 2.4, $E_K(R||f'(A),f'(B)) \neq E_L(R||f'(A),f'(B))$.

Suppose that background knowledge is augmented with the information that A is prior to R, so that SEM L is false. Then the pattern is $B - A \rightarrow R$, and the only two DAGs compatible with the population density and background knowledge are K and M. In that case $E_K(R||f'(A),f'(B)) = E_M(R||f'(A),f'(B))$, and hence the prediction for R is the same for all DAGs compatible with the population density and background knowledge.

If there are many SEMs represented by a pattern, then it may be computationally infeasible to check whether the entailed manipulated expected value is the same in each SEM. In that case a simple modification of the algorithm described in section 6.1 suffices to check this without checking each SEM. Equation 6-1 is still used for $E(R||f'(A),f'(B))$; the only difference is that instead of checking whether $f_K(R|A)$ is invariant under manipulation of A and B, it is also necessary to check whether $f_M(R|A)$ is invariant under manipulation.

Instead of checking each DAG G represented by the pattern individually to determine if $f_G(R|A)$ is invariant under manipulation of A and B, they can all be checked in one step using an augmented pattern (Figure 8). According to the augmented pattern, the only path between A and R is the path $A \rightarrow R$. Hence A is a non-collider on each path from $Policy_A$ and $Policy_B$ to R conditional on A in each DAG represented by the pattern, and the pattern entails that $Policy_A$ and $Policy_B$ are d-separated from R conditional on A in each DAG represented by the pattern. See the Prediction Algorithm (which also handles DAGs that may contain hidden common causes) for more details (Spirtes et al., 2001).

In some cases, different output graphs may predict different values of the target variable. For example, if it is not known that L is false, then $E_K(R||f'(A),f'(B)) \neq E_L(R||f'(A),f'(B))$, even in the large sample limit. (In the augmented pattern in which the edge between A and R is undirected because L is not known to be false, A is a collider on the path from $Policy_A$ to R in some DAGs represented by the pattern, and not others, and hence $f_G(R|A)$ is not invariant under manipulation for each DAG G represented by the pattern.)

There are several possibilities for dealing with this contingency. One possibility is to predict an interval for R, instead of a point value for R; somehow this interval would then have to be used to predict the value of R. Theoretically, an interval could be obtained by calculating $E_G(R||f'(A),f'(B))$ for each SEM G represented by the pattern, and taking the widest limits. Depending upon how many different SEMs there are in the output, this could be computationally infeasible.

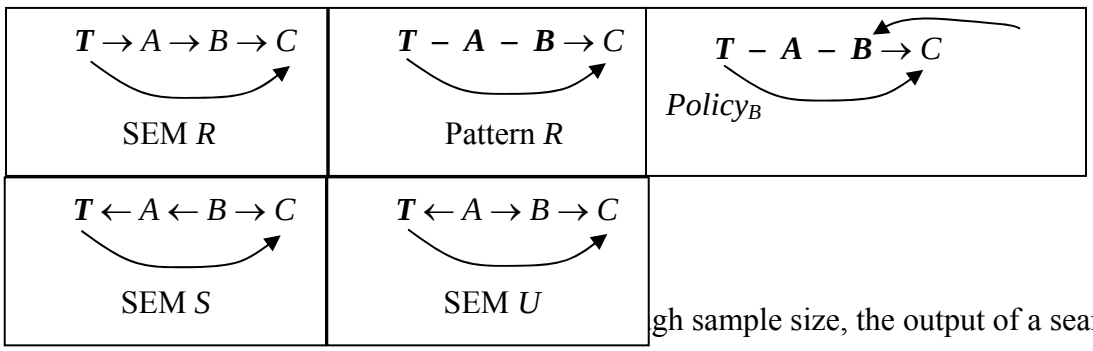
A second possibility is to use a Bayesian approach, and perform model averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999). That is, a prior probability is placed over each G, and a posterior probability for each G is calculated. Then $E_G(R||f'(A),f'(B))$ is

calculated for each G in the output of the search, and the results are averaged together, where each $E_G(R||f'(A),f'(B))$ is weighted by its posterior probability. This requires putting a prior probability over each graph; in addition, if there are many graphs in the output, then this may not be computationally feasible.

6.2.1 Other Examples

In this section, the case where the post-manipulation values of variables other than the target are known is considered. Suppose that SEM R of

Figure 9 is true, and the input is a sample of size n from the unmanipulated population $f(T,A,B,C)$, the post-manipulation values (a, b, c) of A, B, and C respectively, the manipulated density $f'(B) \sim N(0,1)$, and that the target variable is T.



gh sample size, the output of a search will be Pattern R, which represents SEMs R, S, and U. From the Augmented Pattern R, it follows that $Policy_B$ is d-separated from T conditional on A, B, and C in every DAG represented by Pattern R because B is a non-collider on $Policy_B \rightarrow B \rightarrow C \leftarrow T$, and A is a non collider on $Policy_B \rightarrow B - A - T$. Hence $f(T|A,B,C||f'(B)) = f(T|A,B,C)$. This implies that $f(T|A,B,C)$ can be estimated from the unmanipulated population and used to predict T by using the post-manipulation values of A, B, and C. In the context of linear SEMs,

$$\hat{T}(A=a,B=b,C=c) = \hat{E}_{R(\boldsymbol{\theta})}(T \mid A=a,B=b,C=c) =$$
$$E_{R(\boldsymbol{\theta})}(T) + \mathrm{cov}_{R(\boldsymbol{\theta})}(T,\{A,B,C\})\mathrm{var}_{R(\boldsymbol{\theta})}(\{A,B,C\})^{-1} \cdot ([a,b,c] - E_{R(\boldsymbol{\theta})}(\{A,B,C\}))$$

(Instead of $R(\theta)$, S or U could also be used to estimate the various quantities.)

This example can be generalized as follows. Suppose that the input is sample data from the unmanipulated density $f(\mathbf{O},Y)$, the post-manipulation values of $\mathbf{O'} \supseteq \mathbf{MB}(Y,R)$, and neither Y nor a child of Y is manipulated. It follows that $f(Y|\mathbf{O'})$ is invariant under manipulation because the policy variables for the manipulation are d-separated from Y conditional on $\mathbf{O'}$, as the argument in the following paragraph shows. The implication of

this is that the same equation can be used to predict the unmanipulated value of *Y* as is used to predict the manipulated value of *Y*, except that the post-manipulation values of **O'** are used to make the prediction, rather than the unmanipulated values of **O'**. In the particular case of linear SEMs, it is possible to use the following estimated version of Equation 3-1 to predict the value of *Y*:

$$\hat{Y}(\mathbf{O}' = \mathbf{o}') = \hat{E}_{G(\boldsymbol{\theta})}(Y \mid \mathbf{O}' = \mathbf{o}') = E_{G(\boldsymbol{\theta})}(Y) + \mathrm{cov}_{G(\boldsymbol{\theta})}(Y, \mathbf{O}')\mathrm{var}_{G(\boldsymbol{\theta})}(\mathbf{O}')^{-1} \cdot (\mathbf{o}' - E_{G(\boldsymbol{\theta})}(\mathbf{O}'))$$

where $G(\theta)$ is a SEM constructed and estimated from the sample unmanipulated data, but **o'** is the set of post-manipulation values of **O'**. Recall that the middle term on the right hand side is the formula for the partial regression coefficients, so this formula in effect uses the partial regression coefficients of *T* when regressed on *A*, *B*, and *C* in the unmanipulated population.

The reason that the policy variables are d-separated from *Y* conditional on **O'** can be broken into cases. By assumption, $Policy_O$ is neither a parent of *Y* nor a parent of a child of *Y*. If *U* is into *Y* (i.e. *Y* is the tail of an edge on *U*) then some parent of *Y* is a non-collider on *U* and in the conditioning set, so *U* does not d-connect *Y* and $Policy_O$ conditional on **O'**. If *U* is out of *Y*, then there are two cases. First, if some child of *Y* is a non-collider on *U*, then that child is in **MB**(*Y*,*G*) and *U* does not d-connect *Y* and $Policy_O$ conditional on **O'**. If some child of *Y* is a collider on *U*, then some parent of that child is a non-collider on *U* and in **MB**(*Y*,*G*), so *U* does not d-connect *Y* and $Policy_O$ conditional on **O'**. Hence no path *U* d-connects $Policy_O$ and *Y* conditional on **O'**.

If the output of the search is a pattern, then it is always possible to determine from the pattern which variables are in the Markov Blanket of a target. Parents and children of the target will be adjacent to the target in the pattern. Any parent of a child of the target that is itself neither a parent nor a child of the target will be the parent of a child in the pattern (because it forms an unshielded collider in the DAG). For example, from Pattern *R*, the Markov Blanket of *T* is *A*, *B*, and *C*.

However, sometimes it is not possible to determine from a pattern whether a variable in the Markov Blanket is a parent or a child. For example, it is not possible to determine from Pattern *R* whether *A* is a parent or a child of *T*. If *A* is manipulated, then some pair of *R*, *S*, and *U* do not agree on the manipulated density of *T* regardless of what set of variables (including the empty set) is conditioned on.

When a variable that is known to be a child of the target is manipulated, the following line of reasoning is superficially plausible, but fallacious. Suppose that it is known that R is the true SEM, and A is manipulated. After the manipulation, which breaks the edge into A, the Markov Blanket of T is $\{B, C\}$, and A is independent of T. It is tempting to suppose that the partial regression coefficient for A could be set to zero, and the partial regression coefficients $r_{T,B.\{B,C\}}$ and $r_{T,B.\{B,C\}}$ from the unmanipulated population could still be used to predict T.

$$\hat{T}(B=b,C=c)=\hat{E}(T \mid B=b,C=c)=$$
$$E_{R(\boldsymbol{\theta})}(Y)+\mathrm{cov}_{R(\boldsymbol{\theta})}(T,\{B,C\})\mathrm{var}_{R(\boldsymbol{\theta})}(\{B,C\})^{-1}\cdot([b,c]-E_{R(\boldsymbol{\theta})}(\{B,C\}))$$

However, this is incorrect, because $f(Y|B,C)$ is not invariant under manipulation of A if R is the true SEM, and so neither are the partial regression coefficients. The correct method, as described in section 2.4 is to form the new model $R(\theta')$, where θ' sets the coefficient of the edge from T to A to zero, and gives the new density to the error term of A. Then

$$\hat{T}(B=b,C=c)=\hat{E}(T \mid B=b,C=c)=$$
$$E_{R(\boldsymbol{\theta}')}(Y)+\mathrm{cov}_{R(\boldsymbol{\theta}')}(T,\{B,C\})\mathrm{var}_{R(\boldsymbol{\theta}')}(\{B,C\})^{-1}\cdot([b,c]-E_{R(\boldsymbol{\theta}')}(\{B,C\}))$$

7. Causal Inference From Samples

The procedure for inferring the effects of manipulations from randomly drawn independent samples is quite similar to the process of inferring the effects of manipulations from a given population density.

Problem 4: Statistical causal predictive modeling

Input: A sample of n $\{\mathbf{o}, y\}$ examples drawn randomly and independently according to $f(\mathbf{O}, Y)$, a manipulation $f'(X_i=x_i)$ of X_i, the post-manipulation values $\mathbf{o'}$ of a subset $\mathbf{O'}$ of the variables $\mathbf{O}$, and a target variable Y whose post-manipulation value is to be predicted.

Output: $\hat{Y}$ ($\mathbf{O'}=\mathbf{o'}$, $f'(X_i=x_i)$), a predictor of the post-manipulation value of Y that has a small MSPE.

Step 1: A search procedure, such as PC, that is based on conditional independence constraints can use statistical tests, instead of an oracle about conditional independence in the population density, to decide whether a conditional independence constraint holds. As long as the probabilities of type I and type II error go to zero as the sample size approaches infinity (pointwise consistency), the probability of inferring the wrong pattern or DAG also goes to zero, as does the MSPE.

For example, if K is the true SEM and L is known to be false, then at a large enough sample size, the output pattern is $B - A \rightarrow R$. Other searches that have advantages for large number of variables or small samples sizes include Hiton (Aliferis, Tsamardinos, & Statnikov, 2003), HUGIN (Madsen, Lang, Kjaerulff, & Jensen, 2003), PCX (Bai, Glymour, Padman, Ramsey, & Spirtes, 2004). Links to a number of algorithms are collected at http://www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html.

Step 2: After finding a pattern or DAG as the output of a search, step 2 proceeds in exactly the same way, and

$$E_M(R \,\|\, f'(A), f'(B)) = \int_R R \cdot \int_A f(R \mid A) f'(A) dR$$

Step 3: The quantities that are substituted into the formula for the manipulated quantity as a function of the population density can be estimated from the sample. There are several methods of doing this. In one method, the free parameters of the SEM are first estimated, and then those estimates are plugged into a formula for the conditional probability.

For example, choose M ($B \rightarrow A \rightarrow R$). In M, the conditional density $f_M(R|A=a)$ has $\sigma^2_M(R|A) = \sigma^2(\varepsilon_A)$, and $E_M(R|A=a) = b_{R,A} \cdot a$. $\sigma^2(\varepsilon_A)$ and $b_{R,A}$ are both free parameters of M that can be consistently estimated by standard methods as $\hat{\sigma}^2(\varepsilon_A)$ and $\hat{b}_{R,A}$ respectively. Hence $\hat{f}_M(R \mid A = a) = N(\hat{b}_{R,A} \cdot a, \hat{\sigma}^2(\varepsilon_A))$. Then $\hat{f}_M(R \mid A)$ can be calculated from the estimated of the free parameters, and substituted into the formula, producing a consistent estimate of $E_M(R\|f'(A),f'(B))$.

$$E_M(R \,\|\, f'(A), f'(B)) = \int_R R \cdot \int_A \hat{f}_M(R \mid A) f'(A) dR dA$$

It is not always the case that every free parameter of the SEM has to be estimated in order to estimate the quantities in the equation for the manipulated expected value. For

example, $\hat{f}_M(R \mid A)$ depends only upon $b_{R,A}$ and $\sigma^2(\varepsilon_A)$, and not upon $b_{B,A}$. In other cases, it is possible to directly estimate a function of the free parameters without first estimating the free parameters themselves. (This is important for the case where some variables are unobserved, and not all of the free parameters can be consistently estimated.)

7.1.1 Score Based Searches

Given samples rather than a population density, step 1 can also use a score based search. However, instead of using AIC to score models, a number of searches employ the Bayesian Information Criterion (BIC).

$$\text{BIC}(M\text{, sample}) = -2 \cdot L(\Sigma_{M(\hat{\boldsymbol{\theta}})}, \text{sample}) + \ln(n) \cdot \text{df}_M,$$

BIC differs from AIC in that the penalty for complexity in BIC increases with sample size, while the penalty for complexity in AIC does not.

In the large sample limit, each SEM that can represent the density and is simplest (in terms of dimensionality) receives a higher BIC score than any SEM that cannot represent the density, or is more complicated; this is not true of the model with the highest AIC score. For example, if some population density represented by *K* and no SEM with a subgraph of *K*, then no other SEM will receive a higher score than *K* (although *L* and *M* will receive equal scores). As the sample size approaches infinity, the difference between a SEM containing the population density and any other SEM that either does not contain the population density or is not the simplest, approaches infinity. For example, although *O* also contains the population density, at a large enough sample size, *K* (and *L* and *M*) will receive a lower BIC score.

Selecting a true SEM according to the lowest BIC score is only correct if assumptions relating probability densities and causal models are made. If *K* can represent a population probability density, without further assumptions, *O* may be the true SEM even though it is more complex than *K*, and in the large sample limit *K* receives a lower BIC score.

Using simplicity to select a causal model has a different motivation than using simplicity to select a statistical model. In the latter case, the motivation is that simpler models may have smaller MSPE (even if they do not contain the population density); this can be the case even without making any empirical assumptions about causal structures in the world being simple. In contrast, using simplicity to select causal models in effect is

making an empirical assumption that simpler causal models are more likely to be true, other things being equal.

The Causal Faithfulness Assumption (section 5.1.2) is a kind of simplicity assumption. Given the Causal Markov and Faithfulness Assumptions, in the large sample limit, no SEM receives a higher BIC score than the true SEM (Chickering, 2003).[21]

For SEMs that have no unobserved common causes, the Greedy Equivalence Search (Chickering, 2003) returns a Markov equivalence of SEMs highest scoring pattern in the large sample limit with probability 1 (under the Causal Markov and Faithfulness Assumptions). Although in the worst case the algorithm is exponential in the number of variables, in practice it can be used on graphs with relatively few edges but hundreds of variables. The LiNGAM algorithm (Shimizu et al., 2006) is a score-based search that returns a single DAG for linear SEMs with at most one Gaussian error term. The Sparse Candidate Algorithm (Friedman, Nachman, & Pe'er, 1999) is also designed to work on datasets with large numbers of variables.

It is also possible to use a Bayesian approach to modeling to score SEMs, using the posterior probability of a SEM conditional on the data (Chickering, 2003; Heckerman, 1998; Geiger & Heckerman, 1994). This is closely related to using BIC as a score, because under a wide variety of priors, in the large sample limit the posterior probability of a SEM and the BIC score of a SEM favor the same SEMs. However, the Bayesian approach allows the added flexibility of incorporating prior degrees of belief in different graphs and different parameters, which can be important for small sample sizes. In addition to selecting a single SEMs to estimate the value of a variable, the Bayesian approach also value of a variable to be estimated by averaging together the estimates of different SEMs weighted by their posterior probabilities (Hoeting et al., 1999).

8. Unobserved Variables

In contrast to the case of estimating conditional expected values, introducing unobserved variables may be required (in the absence of correlated errors and double-headed arrows in the graph) in order to correctly represent both the population density

[21] In addition, the true distribution is in one of the models, each model has equal prior probability, for a wide variety or priors the difference in BIC scores to between two models M_1 and M_2 is an approximation

and the post-manipulation densities. If SEM *K* is the correct causal model, and *A* is not measured, then *B* and *R* are correlated, but manipulating *B* produces no change in *R*, and manipulating *R* produces no change in *B*. The only linear SEM with a DAG that can represent both the population and the manipulated densities is one in which there is an unmeasured common cause of *B* and *R*. In the case of predicting unmanipulated values of variables, at a given sample size a SEM with unobserved variables may provide predictions with a smaller MSPE than any SEM without any unobserved variables, but in the large sample limit the MSPE approaches a minimum for some SEMs without unobserved variables. In contrast, in the case of predicting manipulated values of variables, in the large sample limit, a SEM with unobserved variables may be the only kind of SEM that has a MSPE that approaches to the minimum.

An alternative to introducing unobserved variables into a model is introducing correlated errors between observed variables. Introducing correlated errors instead of unobserved variables can avoid a number of the statistical problems caused by introducing unobserved variables, and makes the search space finite (Richardson & Spirtes, 2002a; Spirtes et al., 1995). Hence, unless background knowledge rules out the possibility of unobserved common causes, accurate prediction of manipulated variables requires allowing the possibility of unobserved common causes. The possibility of unobserved common causes requires making several modification to the algorithm described in sections 5 and 6.

8.1. Scoring and Searching

8.1.1 Constraint Based Search

For SEMs, that may contain unobserved common causes, there is a constraint based search algorithm (Fast Causal Inference, (Spirtes et al., 1995)) that is a close analog to constraint-based searches over SEMs that do not contain unobserved common causes.

Suppose that the variable *H* in SEM *D* of

---

to the posterior probabilities log $P(M_1|\text{sample})/P(M_2|\text{sample})$.

Figure 7 is an unmeasured variable. *D* does not entail the same set of conditional independence relations as SEM *C* does because it has a different variable set than *C*. For example, $I_C(X_1,X_2|H)$, but it is not the case that $I_D(X_1,X_2|H)$ because *D* does not contain variable *H*. However, if the observed variables O = $\{X_1, X_2, X_3, X_4\}$ then all of the conditional independencies among variables in O that are entailed by *C* (the empty set in this case) are also entailed by *D*, and vice-versa. SEMs *C* and *D* are *Markov equivalent over the set of observed variables O* if they entail the same set of conditional independence relations among the variables in O. The Markov equivalence class over O contains an infinite number of SEMs, because adding more and more unobserved common causes of X_1 and X_2 does not change the conditional independence relations entailed over O.

A partial ancestral graph (PAG) is the analog of a pattern for SEMs that may contain unobserved common causes. Like a pattern, it is a graph that contains information about the ancestor relations common to all SEMs in a given Markov equivalence class over **O** (Spirtes et al., 1995). However, a PAG may contain double-headed edges, indicating that a set of conditional independence relations can only be entailed by a SEM with unobserved common causes (and other kinds of edges as well).

The Fast Causal Inference (FCI) algorithm (Spirtes et al., 2001; Spirtes et al., 1995; Zhang, 2007) performs a series of conditional independence tests and constructs a PAG on the basis of those tests. In the large sample limit, it returns a PAG that contains the true SEM with probability 1 under the Causal Markov and Causal Faithfulness Assumptions. The FCI algorithm is slower than the PC algorithm, and less informative because it typically returns a larger class of candidate SEMs.[22]

In contrast to a score based search, a constraint based search does not require the estimation of parameters, or the calculation of the dimensionality of the marginal density - it requires only being able to perform the appropriate tests of the constraints (e.g. conditional independence constraints) used in the search. Furthermore, although the number of DAGs with unobserved common causes is infinite, the number of Markov equivalence classes over **O** is finite.

---

[22] In some cases, there are no SEMs without unobserved common causes compatible with the pattern of conditional independence constraints that hold in the population.

Conditional independence constraints in acyclic SEMs are examples of non-parametric constraints - they hold for all families of densities. Using non conditional independence constraints for model selection allows a finer discrimination than a Markov equivalence class over **O** for those density families in which the Markov equivalence class over **O** is not the same as a density equivalence class over **O**. Using non conditional independence constraints is particularly important for causal models with hidden common causes, where in many cases there are no conditional independence relations among the variables in **O**.

For example, SEMs *C* and *D* are Markov equivalent over **O**, but they are not distribution equivalent over **O**. *C*, unlike *D*, entails "vanishing tetrad constraints", i.e. $\mathrm{cov}_C(X_1,X_2) \cdot \mathrm{cov}_C(X_3,X_4) = \mathrm{cov}_C(X_1,X_3) \cdot \mathrm{cov}_C(X_2,X_4) = \mathrm{cov}_C(X_1,X_4) \cdot \mathrm{cov}_C(X_2,X_3)$ (Silva, Scheines, Glymour, & Spirtes, 2006; Harman, 1976). This is a constraint on the marginal density over **O** that is not a conditional independence constraint. So when unobserved common causes are allowed, SEMs that are Markov equivalent (over **O**) are not necessarily distribution equivalent over **O** even when all of the error terms are Gaussian. If the vanishing tetrad constraint were false, then it would be known that *C* is not the true SEM, even though it does not entail any false conditional independence constraints among the observed variables. While there are recently developed algorithms for determining when two SEMs are distribution equivalent over **O**, they are so computationally intensive they are only practical for SEMs with a few variables (Geiger & Meek, 1999).

When Markov equivalence fails to entail density equivalence, then using conditional independence relations alone for causal inference is still correct, but it is not as informative as theoretically possible. For example, assuming causal sufficiency and non-Gaussian errors (Shimizu et al., 2006), conditional independence tests can at best reliably determine the correct Markov equivalence class, while using other features of the sample density can be used to reliably determine a unique graph (Shimizu et al., 2006). The situation is similar for Gaussian errors where there are unobserved common causes, which may entail vanishing tetrad constraints, or other kinds of constraints. The difficulty is that in many cases it is not known how to use the extra information contained in the

density to reliably narrow down the set of DAGs output. However, in some cases it is known how to use the extra information (Silva et al., 2006).

8.1.2 Score-Bases Search and BIC

Using a score such as BIC to directly search over the space of SEMs that may contain unmeasured common causes raises a number of extra difficulties. The space of SEMs that may contain unobserved common causes is infinite, and it is not clear how to order the SEMs to be searched.

Furthermore, while the use of penalized likelihood scores, such as BIC, for SEMs with no unobserved common causes is not problematic, there are major statistical problems in scoring models using penalized likelihood scores for SEMs with unobserved common causes. In order to calculate a BIC score it is necessary to calculate a maximum likelihood estimate of the SEM parameters, and the dimensionality of the set of marginal densities over the observed variables represented by the SEM.

However, with the exception of a few densities such as Gaussian, or multinomial, even when the joint density falls into a family of densities that is well understood, the marginal densities will not. In cases where the marginal density is a member of a well understood family of densities, the parameters of the SEM may not be identifiable at all, and hence it is not possible to find a maximum likelihood estimate of the parameters from the data over the observed marginal. Furthermore, even in those cases where the maximum likelihood estimates can be calculated (such as C where H is unobserved) the actual calculations typically involve an iterative hill-climbing algorithm that is much more computationally expensive than regression and can get stuck in local maxima.

In addition, there are both theoretical and practical difficulties in calculating the dimensionality of the marginal probability density that are represented by a SEM with unobserved common causes. The dimensionality is not well defined for some values of the parameters, and is difficult to calculate even when it is well defined (Geiger et al., 2002).

This problem is caused by the fact that while the unobserved conditional independence relations (those that involve the unobserved common cause) cannot be tested directly, they can nevertheless entail constraints on the marginal density that are not conditional independence relations. These non-conditional independence constraints

present both a problem and an opportunity. On the one hand, if it is known how to use them to eliminate some graphs from consideration (as in certain special cases) then they strengthen the causal inferences that can be made. On the other hand, the non-conditional independence constraints are the reason that the maximum likelihood estimate can get stuck in local maxima and that the dimensionality of the marginal densities represented by a SEM with unobserved common causes is sometimes undefined or difficult to calculate.

There has been some progress towards generalizing a GES style search to cover SEMs with unobserved common causes. In the Gaussian case , it is possible to parameterize PAGs in such a ways that they impose only the conditional independence constraints entailed by the SEMs they represent. There are a finite number of these kinds of graphical models for a given set of observed variables, and it is known how to calculate the maximum likelihoods and dimensions needed to calculate BIC scores for these models (Richardson & Spirtes, 2002a). In addition, score-based searches for limited numbers of hidden common causes have been proposed by Friedman (Friedman, 1998), and Heckerman (Heckerman, 1998).

8.2. Calculating Manipulated Densities from Causal Graphs

If some of the variables are unmeasured, and only the marginal density over the observed set of variables **O** is given, in some cases the effect of a manipulation is not a function of the marginal density. In other cases the effect of a manipulation is a function of the marginal density, but may not be equal to any partial regression coefficient, no matter what set of (unmanipulated) observed variables is regressed on

Given a single DAG with hidden common causes, Pearl's do-calculus extended the necessary and sufficient conditions for determining which conditional probabilities were invariant to sequences of manipulations, and showed how a much broader range of manipulated quantities could be expressed in terms of invariant quantities (Pearl, 1995). More recently, complete algorithms have been developed for finding formulas that express various kinds of manipulated quantity in terms of observed invariant marginal densities (Shpitser & Pearl, 2006a; Shpitser & Pearl, 2006b; Huang & Valtorta, 2006).

In many cases, without strong background knowledge, the output of a search for causal models is a Markov equivalence class over **O** or a density equivalence class over **O**. In those cases, the problem of finding the effects of a manipulation can only be done for those cases in which each member of the equivalence class agrees on the effects of the manipulation. Sufficient conditions for determining when a given conditional probability is invariant under a manipulation have been developed, and these can be used to express some manipulated quantities as functions of conditional probabilities that are invariant under manipulation, regardless of which of the SEMs in the Markov equivalence class over **O** are true (Spirtes et al., 1993). The Prediction Algorithm generalizes the algorithm for calculating manipulated conditional densities described in 6.2 to SEMs with hidden common causes. It uses d-separation from a policy variable to determine when conditional probabilities are invariant under manipulation in PAGs instead of patterns (Spirtes et al., 1993; Spirtes et al., 2001).

One important limitation on reasoning about SEMs with hidden common causes, is that it is not the case that under the Causal Markov and Faithfulness assumptions although it is possible to consistently estimate conditional expected values in the large simple limit (in the sense that the mean squared error of the estimate approaches zero), it is not possible to put non-trivial probabilistic bounds on the mean square prediction error at any finite sample size, which would require uniform consistency ((Robins, Scheines, Spirtes, & Wasserman, 2003).

9. Summary

The following is a brief summary of some important differences between the problem of predicting the value of an variable in an unmanipulated population from a sample, and the problem of predicting the post-manipulation value of a variable from a sample from an unmanipulated population. In an unmanipulated population *P*, the predictor that minimizes the MSPE is the conditional expected value.

1. $E(Y|\mathbf{O})$ is a function of $f(\mathbf{O},Y)$, regardless of what the true causal model is.[23] Given $f(\mathbf{O},Y)$, in order to calculate $E(Y|\mathbf{O})$ there is no need to do feature selection, or search for a statistical model. In contrast, $E_P(Y|\mathbf{O'}||f'(O))$ is a

function of $f(\mathbf{O},Y)$ and at least some features of the causal relations between variables. In some cases, if G and H are in the set of causal models compatible with the $f(\mathbf{O},Y)$ and background knowledge, $E_G(Y|\mathbf{O'}||f'(O)) \neq E_H(Y|\mathbf{O'}||f'(O))$, in which case the conditional expected values in the manipulated population is not a function of the background knowledge and $f(\mathbf{O},Y)$. In other cases $E_G(Y|\mathbf{O'}||f'(O)) = E_H(Y|\mathbf{O'}||f'(O))$ for all causal models G and H compatible with $f(\mathbf{O},Y)$ and background knowledge, in which case $E_P(Y|\mathbf{O'}||f'(O))$ is a function of the background knowledge and the $f(\mathbf{O},Y)$.

2. In order to determine whether $E_P(Y|\mathbf{O'}||f'(O))$ is a function of $f(\mathbf{O},Y)$ and background knowledge, it is necessary to find all of the causal models compatible with $f(\mathbf{O},Y)$ and background knowledge, not simply one causal model compatible with $f(\mathbf{O},Y)$ and background knowledge.
3. Determining which causal models are compatible with background knowledge and a $f(\mathbf{O},Y)$ requires making additional assumptions connecting population densities to causal models (e.g. Causal Markov and Faithfulness).
4. Without introducing some simplicity assumptions about causal models, for some common families of densities (e.g. Gaussian, multinomial), no $E_P(Y|\mathbf{O'}||f'(O))$ are functions of the population density without very strong background knowledge.
5. The justification for using simple statistical models is fundamentally different than the justification for using simple causal models. At a given sample size, the use of simple statistical model can be justified even if causal relations are not simple. However, the assumption that the simplest causal model compatible with $f(\mathbf{O},Y)$ and background knowledge is a substantive assumption about the simplicity of mechanisms that exist in the world.
6. For many families of densities (e.g. Gaussian, multinomial), there is always a statistical model without hidden variables that contains the population density. For those same families of densities, a causal model that contains both the population probability density and the post-manipulation probability densities may require the introduction of unobserved variables.

---

[23] This ignores the problem of conditioning on sets of measure zero.

7. Given a population density, and the set of causal models consistent with the population density and background knowledge, calculating the effects of a manipulation can be difficult because:
   a. There may be unobserved variables (even if only a single causal model is consistent with $f(\mathbf{O},Y)$ and background knowledge).
   b. There may be multiple causal models compatible with $f(\mathbf{O},Y)$ and background knowledge.
8. For non-experimental data, a post-manipulation density is different than the population density from which the sample is drawn. The post-manipulation values of the target variable Y are not directly measured in the sample. Hence, it is not possible to estimate the error in $E_P(Y|\mathbf{O'}||f'(O))$ by comparing it to the values in a sample from the $f(\mathbf{O},Y)$.

10. Open Problems[24]

Each of the elements used to solve the simple examples in this tutorial raise a number of questions. The questions listed below are areas of active research that have produced some answers, but many open questions remain. The proceedings of the *Uncertainty and Artificial Intelligence* conferences and the *Statistics and Artificial Intelligence* conferences contain many articles related to these areas of research. In addition this Special Issue on Causation of the Journal of Machine Learning Research contains articles relevant to these questions, as explained in the tutorial. The citations given are just a sample from larger bodies of research.

10.1. Models

There are a wide variety of causal models that have been employed in different disciplines. These include Bayesian Networks, Chain Graphs, Partial Ancestral Graphs, Markov Decision Processes, Structural Equation Models, Propensity Scoring, Information Theory, and Granger Causality. The relative advantages and disadvantages of these models and the relationships between these models are partially, but not

[24] The content and organization of this section are largely due to suggestions from Constantin Aliferis, whom I thank for his suggestions.

completely understood. What new models are appropriate for different domains, e.g. feedback or reversible systems (Spirtes et al., 1993; Spirtes et al., 2001)? What new models are appropriate for different combinations of kinds of data, e.g. experimental and observational (Eberhardt & Clark Glymour, 2006; Eberhardt et al., 2005; Yoo & Cooper, 2004; Yoo, Cooper, & Schmidt, 2006; Danks, 2002; Cooper & Yoo, 1999)? What new models are appropriate for different kinds of background knowledge, and different families of densities?

10.2. Model Scores

What kind of scores can be used to best evaluate causal models from various kinds of data? While some scores, such as BIC, have good large sample properties, they are difficult to compute or cannot be applied to some causal models, and may not have good small sample properties. In a related vein, what are good families of prior distributions that capture various kinds of background knowledge?

10.3. Search Algorithms

How can search algorithms be improved to incorporate different kinds of background knowledge, search over different classes of causal models, run faster, handle more variables and larger sample sizes, be more reliable at small sample sizes, and produce output that is as informative as possible?

10.4. Properties of Search Algorithms

For existing and novel causal search algorithms, what are their semantic and syntactic properties (e.g. soundness, consistency, maximum informativeness)? What are their statistical properties (pointwise consistency, uniform consistency, sample efficiency)? What are their computational properties (computational complexity)?

10.5. Assumptions

What plausible alternatives are there to the Causal Markov and Faithfulness Assumptions? Are there other assumptions might be weaker and hold in more domains and applications without much loss about what can be reliably inferred? Are there stronger assumptions that are plausible for some domains that might allow for stronger

causal inferences? How often are these assumption violated, and how much do violations of these assumptions lead to incorrect inferences?

There are special assumptions, such as linearity, which can improve the strength of causal conclusions that can be reliably inferred, and the speed and sample efficiency of algorithms that draw the conclusions. What other distribution families or stronger assumptions about a domain are there that are plausible for some domains and how can they be used to improve causal inference?

Can various statistical assumptions be relaxed? For example, what if the sample selection process is not i.i.d., but may be causally affected by variables of interest (Richardson & Spirtes, 2002a; Spirtes et al., 1995; Cooper, 1995; Cox & Wermuth, 1996a; Cooper, 2000)?

10.6. Deriving Consequences of Causal Models

Shpitser and Pearl have given complete algorithms for deriving the consequences of various causal models with hidden common causes in terms of the unmanipulated density and the given manipulation (Shpitser & Pearl, forthcoming). Partial extensions of these results to deriving consequences from sets of causal models have been given (Zhang, forthcoming); are there further extensions to derivations from sets of causal models?

It is often useful to quickly derive constraints (e.g. vanishing tetrad constraints) on marginal densities from causal models with hidden common causes, in order to guide search. Are there other constraints on densities that can be derived from causal models, and how can they be incorporated into search algorithms?

10.7. Applications

Application of causal inference algorithms to many domains would be highly useful both in terms of testing and improving causal inference algorithms, suggesting new problems, and in producing domain knowledge.

10.8. Evaluation

What are the most appropriate performance measures for causal inference algorithms? What benchmarks can be established? What is the best research design for testing causal inference algorithms?

10.9. Interconnections

Many different domains have studied causal discovery including Artificial Intelligence, Econometrics, Markov Decision Processes, Operations Research, Control Theory, Experimental Design, and Statistics. What are the formal connections between the different models, assumptions, and algorithms used in each of these domains? What can each of these domains learn from the others?

11. Appendix

11.1. Estimators

An estimator ϕ_n of a model parameter μ (such as the expected value of a variable) is a function from samples O^n of size n to a real number - that is, for each sample, the estimator outputs an estimate of the quantity μ. The quality of an estimator ϕ_n of μ can be measured by its *mean square error*: that is, the expected value (over all randomly selected samples of size n) of $(\phi_n(O^n) - \mu)^2$, the square of the difference between the real number output by the estimator and μ. Let $\bar{\phi}_n$ be the average output of $\phi_n(O^n)$ (with respect to the sampling density of O^n). The expected value squared error is the sum of two terms: the bias and the variance of the estimator. The *bias* of the estimator ($\bar{\phi}_n - \mu$), i.e. the difference between the expected value output of the estimator and the true value μ. The *variance* of the estimator is the expected value of $(\phi_n(O^n) - \bar{\phi}_n))^2$, i.e. the expected value of the squared difference between the output of the estimator and the expected value of the estimator). A *pointwise consistent estimator* is one in which the mean squared error approaches zero as the sample size approaches infinity. The quality of an estimator at a finite sample size depends upon the two factors of the bias of an estimator and the variance of the estimator. (There are other desirable properties that

estimators can have, such as being *uniformly consistent* - roughly that it is possible to put probabilistic bounds on the size of the error at a given sample size - that will not be discussed here. (See Robins et al., 2003).)

One kind of estimator that is commonly employed is a maximum likelihood estimator, which under mild regularity conditions has a number of desirable properties such as pointwise consistency. Suppose that SEM *C* is given and the goal is to estimate the values of the free parameters of *C*. For assignment θ of values to the free parameters of *C*, there is an implied covariance matrix. Given the implied covariance matrix, it is possible to determine the probability (density) of drawing sample data that has the observed sample covariance matrix; this is the likelihood of the data for θ. A maximum likelihood estimator selects the assignment θ of values to the free parameters that makes the sample data have highest value for the probability density. In the case of a SEM over a causally sufficient set of variables, the maximum likelihood estimate of the linear coefficient $b_{R,A}$ (denoted by $\hat{b}_{R,A}$) is the regression coefficient of *A* when *R* is regressed on its non-descendants in the causal graph (e.g. in the case of SEM *K*, the regression coefficient of *R* when *A* is regressed on $\{A, B\}$). In *K*, the formula for the total effect of *A* on *R* is $m \cdot b_{R,A}$. Substituting the maximum likelihood estimates of $\hat{b}_{R,A}$ into the formula yields $m \cdot \hat{b}_{R,A}$, which is a maximum likelihood estimate of the total effect of *A* on *R*.

11.2. Graph Terminology

For a directed edge $A \rightarrow B$, *A* is the *tail* of the edge and *B* is the *head* of the edge, *A* is a *parent* of *B*, and *B* is a *child* of *A*. **Parents**(*A*,*G*) is the set of parents of *A* in *G*. A *path U* between X_a and X_b is a sequence of edges $<E_1,...,E_m>$ such that one endpoint of E_1 is X_a, one endpoint of E_m is X_b, and for each pair of consecutive edges E_i, E_{i+1} in the sequence, $E_i \neq E_{i+1}$, and one endpoint of E_i equals one endpoint of E_{i+1}. A *directed path P* between X_a and X_b is a sequence of directed edges $<E_1,...,E_m>$ such that the tail of E_a is X_1, the head of E_m is X_b, and for each pair of edges E_i, E_{i+1} adjacent in the sequence, $E_i \neq E_{i+1}$, and the head of E_i is the tail of E_{i+1}. A *vertex occurs on a path* if it is an endpoint of one of the edges in the path. A path is *acyclic* if no vertex occurs more than once on the path. A vertex *A* is an *ancestor* of *B* (and *B* is a *descendant* of *A*) if and only if either

there is a directed path from A to B or $A = B$. **ND**(A,G) is the set of non-descendants of A in G. A vertex X is a *collider* on undirected path U if and only if U contains a subpath $Y \rightarrow X \leftarrow Z$; otherwise if X is on U it is a *non-collider* on U. X is an *ancestor of a set* of vertices **Z** if X is an ancestor of some member of **Z**.

12. Bibliography