GRAPHS, CAUSALITY, AND STRUCTURAL EQUATION MODELS

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

Abstract

Structural equation modeling (SEM) has dominated causal analysis in the social and behavioral sciences since the 1960s. Currently, many SEM practitioners are having difficulty articulating the causal content of SEM and are seeking foundational answers. Recent developments in the areas of graphical models and the logic of causality show potential for alleviating such difficulties and thus for revitalizing structural equations as the primary language of causal modeling. This paper summarizes several of these developments, including the prediction of vanishing partial correlations, model testing, model equivalence, parametric and nonparametric identifiability, control of confounding, and covariate selection. These developments clarify the causal and statistical components of structural equation models and the role of SEM in the empirical sciences.

1 INTRODUCTION

1.1 Causality in Search of a Language

The word *cause* is not in the vocabulary of standard probability theory. It is an embarrassing yet inescapable fact that probability theory, the official mathematical language of many empirical sciences, does not permit us to express sentences such as "Mud does not cause rain"; all we can say is that the two events are mutually correlated, or dependent – meaning that if we find one, we can expect to encounter the other. Scientists seeking causal explanations for complex phenomena or rationales for policy decisions must therefore supplement the language of probability with a vocabulary for causality, one in which the symbolic representation for the causal relationship "Mud does not cause rain" is distinct from the symbolic representation for "Mud is independent of rain." Oddly, such distinctions have not yet been incorporated into standard scientific analysis.[1]

[1] A summary of attempts by philosophers to reduce causality to probabilities is given in [Pearl, 1996, pp. 396–405].

Two languages for causality have been proposed: path analysis or structural equation modeling (SEM) [Wright, 1921; Haavelmo, 1943], and Neyman-Rubin's potential-response model [Neyman, 1923; Rubin, 1974]. The former has been adopted by economists and social scientists [Goldberger, 1972; Duncan, 1975], while a small group of statisticians champion the latter [Rubin, 1974; Robins, 1986; Holland, 1988]. These two languages are mathematically equivalent,[2] yet neither has become standard in causal modeling – the structural equation framework because it has been greatly misused and inadequately formalized [Freedman, 1987], and the potential-response framework because it has been only partially formalized and, more significant, because it rests on an esoteric and seemingly metaphysical vocabulary of counterfactual variables that bears no apparent relation to ordinary understanding of cause-effect processes.

Currently, potential-response models are understood by few and used by even fewer, while structural equation models are used by many but their causal interpretation is generally questioned or avoided. The main purpose of this paper is to formulate the causal interpretation and outline the proper use of structural equation models, and thus to reinstate confidence in SEM as the primary formal language for causal analysis in the social and behavioral sciences. But first, a brief analysis of the current crisis in SEM research in light of its historical development.

1.2 Causality and Structural Models

SEM was developed by geneticists [Wright, 1921] and economists [Haavelmo, 1943; Koopmans, 1950, 1953] so that qualitative cause-effect information could be combined with statistical data to provide quantitative assessment of cause-effect relationships among variables of interest. Thus, to the often asked question, "Under what conditions can we give causal interpretation to structural coefficients?" Wright and Haavelmo would have answered, "Always!" According to the founding fathers of SEM, the conditions that make the equation $y = \beta x + \epsilon$ *structural* are precisely those that make the causal connection between X and Y have no other value but β and nothing about the statistical relationship between x and ϵ can ever change this interpretation of β. Amazingly, this basic understanding of SEM has all but disappeared from the literature, leaving modern econometricians and social scientists in a quandary over β.

Most SEM researchers today are of the opinion that extra ingredients are necessary for structural equations to qualify as carriers of causal claims. Among social scientists, James, Mulaik, and Brett [1982, p. 45] for example, state that a condition called *self containment* is necessary for consecrating the equation $y = \beta x + \epsilon$ with causal meaning, where self-containment stands for $cov(x, \epsilon) = 0$. According to James, Mulaik and Brett, whenever self-containment does not hold "neither the equation nor the functional relation represents a causal relation." Bollen [1989, p. 44] reiterates the necessity of self containment (under the rubric *isolation* or *pseudo-isolation*), contrary to the understanding that structural equations attain their causal interpretation prior to, and independently of, any statistical relationships among their constituents. Since the early 1980, it has become exceedingly rare to find an

[2] The equivalence of the potential-response and structural equation frameworks, anticipated by Holland [1986], Pratt and Schlaiffer [1988], Pearl [1995], and Robins [1995] is proven formally in [Galles and Pearl, 1998].

open endorsement of the original SEM logic, namely, that ϵ is defined in terms of β, not the other way around, and that the orthogonality condition $cov(x, \epsilon) = 0$ is neither necessary nor sufficient for the causal interpretation of β (see Section 4.1). In fact this condition is not necessary even for the identification of β, once β is interpreted (See the identification of α in Figs. 8 and 10, below).

Econometricians have just as much difficulty with the causal reading of structural parameters. Leamer [1985, p. 258] observes, "It is my surprising conclusion that economists know very well what they mean when they use the words 'exogenous,' 'structural,' and 'causal,' yet no textbook author has written adequate definitions." There has been little change since Leamer made these observations. Hendry [1995, p. 62], for instance, amplifies the necessity of the orthogonality condition, and states: "...the status of β may be unclear until the conditions needed to estimate the postulated model are specified. For example, in the model:

$$y_t = z_t\beta + u_t \text{ where } u_t \sim \text{ IN } [0, \sigma_u^2],$$

until the relationship between z_t and u_t is specified the meaning of β is uncertain since $E[z_t u_t]$ could be either zero or non-zero on the information provided." LeRoy [1995, p. 211] goes even further: "It is a commonplace of elementary instruction in economics that endogenous variables are not generally causally ordered, implying that the question 'What is the effect of y_1 on y_2' where y_1 and y_2 are endogenous variables is generally meaningless." According to LeRoy, causal relationships cannot be attributed to any variable whose causes have separate influence on the effect variable, a position that denies causal reading to most of the structural parameters that economists and social scientists labor to estimate.

Cartwright [1995, p. 49], a renowned philosopher of science, addresses these difficulties by initiating a renewed attack on the tormenting question, "*Why* can we assume that we can read off causes, including causal order, from the parameters in equations whose exogenous variables are uncorrelated?" Cartwright, like SEM's founders, recognizes that causes cannot be derived from statistical or functional relationships alone and that causal assumptions are prerequisite for validating any causal conclusion. Unlike Wright and Haavelmo, however, she launches an all-out search for the assumptions that would endow the parameter β in the regression equation $y = \beta x + \epsilon$ with a legitimate causal meaning and endeavors to prove that the assumptions she proposes are indeed sufficient. What is revealing in Cartwright's analysis is that she does not consider the answer Haavelmo would have provided, namely, that the assumptions needed for drawing causal conclusions from parameters are already encoded in the *syntax* of the equations and can be read off the associated graph as easily as a shopping list[3]; they need not be searched for elsewhere, nor do they require new proofs of sufficiency. Haavelmo's answer applies to models of any size and shape, including models with correlated exogenous variables.

[3]These assumptions are explicated and operationalized in Section 4. Briefly, if G is the graph associated with a causal model that renders a certain parameter identifiable, then two assumptions are sufficient for authenticating the causal reading of that parameter, namely, (1) every missing arrow, say between X and Y, represents the assumption that X has no effect on Y once we intervene and hold certain other variables fixed, and (2) every missing bi-directed arc $X \leftarrow - - \rightarrow Y$ represents the assumption that all omitted factors that affect Y are uncorrelated with those that affect X. Each of these assumptions is *testable* in experimental settings, where interventions are feasible (Section 4.1.3).

These examples partake of an alarming tendency among economists and social scientists to view a structural equation as an algebraic object that carries functional and statistical assumptions but is void of causal content. This statement from one leading social scientist is typical: "It would be very healthy if more researchers abandoned thinking of and using terms such as cause and effect" [Muthen, 1987, p. 180]. Perhaps the boldest expression of this tendency was recently voiced by Holland [1995, p. 54]: "I am speaking, of course, about the equation: $\{y = a + bx + \epsilon\}$. What does it mean? The only meaning I have ever determined for such an equation is that it is a shorthand way of describing the conditional distribution of $\{y\}$ given $\{x\}$."[4]

The founders of SEM had an entirely different conception of structures and models. Wright [1923, p. 240] declared that "prior knowledge of the causal relations is assumed as prerequisite" in the theory of path coefficients, and Haavelmo [1943] explicitly interpreted each structural equation as a statement about a hypothetical controlled experiment. Likewise, Marschak [1950] and Koopmans [1953] stated that the purpose of postulating a structure behind the probability distribution is to cope with the hypothetical changes that can be brought about by policy. One wonders, therefore, what has happened to SEM over the past 50 years, and why the basic (and still valid) teachings of Wright, Haavelmo, Marschak, and Koopmans have been forgotten.

Some economists attribute the decline in the understanding of structural equations to Lucas' critique [Lucas Jr., 1976], according to which economic agents anticipating policy interventions would tend to act contrary to SEM's predictions which often ignore such anticipations. However, since Lucas' critique merely shifts the model's invariants and the burden of structural modeling from the behavioral level to a deeper level, involving agents' motivations and expectations, it does not exonerate economists from defining and representing the causal content of structural equations at some level of discourse.

I believe that the causal content of SEM has gradually escaped the consciousness of SEM practitioners mainly for the following reasons:

1. SEM practitioners have sought to gain respectability for SEM by keeping causal assumptions implicit, since statisticians, the arbiters of respectability, abhor assumptions that are not directly testable.

2. The algebraic language that has dominated SEM lacks the notational facility needed to make causal assumptions, as distinct from statistical assumptions, explicit. By failing to equip causal relations with precise mathematical notation, the founding fathers in fact committed the causal foundations of SEM to oblivion. Their disciples today are seeking foundational answers elsewhere.

Let me elaborate on the latter point. The founders of SEM understood quite well that in structural models the equality sign conveys the asymmetrical relation "is determined by," and hence behaves more like an assignment symbol (:=) in programming languages than like an algebraic equality. However, perhaps for reasons of mathematical purity (i.e., to avoid the appearance of syntax sensitivity), they refrained from introducing a symbol to represent

[4] Holland's interpretation is at variance with the structural reading of the equation [Haavelmo, 1943], which is "In an ideal experiment where we control X to x and any other set Z of variables (not containing X or Y) to z, Y is independent of z and is given by $a + bx + \epsilon$" (see Section 4.1.1).

the asymmetry. According to Epstein [1987], in the 1940s Wright gave a seminar on path coefficients to the Cowles Commission (the breeding ground for SEM), but neither side saw particular merit in the other's methods. Why? After all, a diagram is nothing but a set of nonparametric structural equations in which, to avoid confusion, the equality signs are replaced with arrows.

My explanation is that the early econometricians were extremely careful mathematicians who thought they could keep the mathematics in purely equational-statistical form and just reason about structure in their heads. Indeed, they managed to do so surprisingly well, because they were truly remarkable individuals who *could* do it in their heads. The consequences surfaced in the early 1980s, when their disciples began to mistake the equality sign for an algebraic equality. The upshot was that suddenly the "so-called disturbance terms" did not make any sense at all [Richard, 1980, p. 3]. We are living with the sad end to this tale. By failing to express their insights in mathematical notation, the founders of SEM brought about the current difficulties surrounding the interpretation of structural equations, as summarized by Holland's "What does it mean?"

1.3 Graphs as a Mathematical Language: An Example

Certain recent developments in graphical methods promise to bring causality back into the mainstream of scientific modeling and analysis. These developments involve an improved understanding of the relationships between graphs and probabilities, on the one hand, and graphs and causality, on the other. But the crucial change has been the emergence of graphs as a mathematical language. This mathematical language is not simply a heuristic mnemonic device for displaying algebraic relationships, as in the writings of Blalock [1962] and Duncan [1975]. Rather, graphs provide a fundamental notational system for concepts and relationships that are not easily expressed in the standard mathematical languages of algebraic equations, and probability calculus. Moreover, graphical methods now provide a powerful symbolic machinery for deriving the consequences of causal assumptions when such assumptions are combined with statistical data.

A concrete example that illustrates the power of the graphical language will set the stage for the discussions in Sections 2 and 3. One of the most frustrating problems in causal analysis has been *covariate selection* – for instance, determining whether a variate Z can be added to a regression equation without biasing the result. More generally, whenever we try to evaluate the effect of one factor (X) on another (Y), we wonder whether we should adjust for possible variations in some other variable, Z, sometimes called a *covariate*, *concomitant*, or *confounder*. Adjustment amounts to partitioning the population into groups that are homogeneous relative to Z, assessing the effect of X on Y in each homogeneous group, and, finally, averaging the results.

The elusive nature of such an adjustment was recognized as early as 1899, when Pearson and Yule discovered what is now called *Simpson's paradox*, namely, that any statistical relationship between two variables may be reversed or negated by including additional factors in the analysis. For example, we may find that students who smoke obtain higher grades than those who do not smoke; but after we adjust for age, smokers obtain lower grades than nonsmokers in every age group; but after we further adjust for family income, smokers obtain higher grades than nonsmokers in every income-age group; and so on.[5]

Despite a century of analysis, Simpson's reversal phenomenon continues to "trap the unwary" [Dawid, 1979, p. 5], and the main question – whether an adjustment for a given covariate Z is appropriate in any given study – continues to be decided informally, case-by-case, with the decision resting on folklore and intuition rather than on hard mathematics. The standard statistical literature is remarkably silent on this issue. Aside from noting that one should not adjust for a covariate that is affected by the putative cause (X),[6] it provides no guidelines as to what covariates might be admissible for adjustment and what assumptions would be needed for making such a determination formally. The reason for this silence is clear: the solution to the covariate selection problem rests on causal assumptions, as we shall see in Section 3, and such assumptions cannot be expressed formally in the standard language of statistics.

In the potential-response framework, a criterion called *ignorability* has been advanced to address the covariate selection problem [Rosenbaum and Rubin, 1983]. It states that Z is an admissible covariate relative to the effect of X on Y if, for every x, the value that Y would obtain had X been x is conditionally independent of X, given Z. This criterion paraphrases the problem in the language of counterfactuals without providing a working test for covariate selection. Because counterfactuals are not observable, and judgments about the conditional independence of counterfactuals are not readily made using our ordinary understanding of causal processes, ignorability has remained a theoretical construct with only a minor impact on practice. Epidemiologists, for example, well apprised of ignorability analysis via the admirable papers of Robins [1986] and Greenland and Robins [1986], are still debating the meaning of "confounding" [Grayson, 1987] and often adjust for the wrong sets of covariates [Weinberg, 1993]. Social scientists, likewise, despite the potential-response analyses of Holland and Rubin [1983] and Sobel [1995], are still struggling with various manifestations of the Lord paradox (a version of Simpson's paradox) in psychometric research [Wainer, 1991] and are still not distinguishing collapsibility from nonconfounding [Steyer et al., 1996].

In contrast, formulating the covariate selection problem in the language of graphs immediately yields a general solution that is both natural and formal. The investigator expresses causal knowledge (i.e., assumptions) in the familiar qualitative terminology of path diagrams, and, once the diagram is complete, a simple procedure decides whether a proposed adjustment (or regression) is appropriate relative to the quantity under evaluation. This procedure, called the *back-door criterion* in Section 3 (Theorems 6 and 7) proceeds roughly as follows: to determine whether a set of variables Z should be adjusted for when we wish to evaluate the total effect of X on Y, we delete all arrows emanating from node X and then test whether, in the resulting graph, all paths between X and Y are *blocked* by nodes corresponding to Z. If the direct effect is to be evaluated, then only the arrow from X to Y should be deleted before applying the test (Theorem 6). The notion *blocked* is defined

[5] The classic case demonstrating Simpson's reversal is the study of Berkeley's alleged sex bias in graduate admission [Bickel et al., 1975], where data showed a higher rate of admission for male applicants overall but, when broken down by departments, yielded a slight bias toward female applicants.

[6] This advice, which rests on the causal relationship "not affected by" is, to the best of my knowledge, the *only* causal notion that has found a place in statistics textbooks. The advice is necessary, but it is not sufficient. The other common guideline, that X should not precede Z [Shafer, 1996, p. 326], is neither necessary nor sufficient, as will become clear in Section 3.

formally in Section 2.1.2 (Definition 1).

This example is not an isolated instance of graphical methods affording clarity and understanding. In fact, the conceptual basis for SEM achieves a new level of precision through graphs. What makes a set of equations "structural," what assumptions the authors of such equations should examine, what the testable implications of those assumptions are, and what policy claims a given set of structural equations advertise are some of the questions that receive simple and mathematically precise answers via graphical methods. These and related issues in SEM will be discussed in the following sections.

1.4 Paper Outline

The testable implications of structural models are explicated in Section 2. For recursive models (herein termed *Markovian*), we find that the statistical content of a structural model can be fully characterized by a set of vanishing partial correlations that are entailed by the model. These vanishing partial correlations can be read off the graph using a simple criterion, called *d-separation*, which applies to both linear and nonlinear models (Section 2.1). The application of this criterion to model testing is discussed in Section 2.2. The *d*-separation criterion leads to graphical tests of model equivalence, which, again, apply to both linear and nonlinear models (Section 2.3).

Section 3 deals with the issue of determining the identifiability of structural parameters prior to gathering any data. In Section 3.1, simple graphical tests of identifiability are developed for linear Markovian and semi-Markovian models (i.e., acyclic diagrams with correlated errors). Extensions to nonparametric models are developed in Sections 3.2 and 3.3, and their ramifications for practical problems of covariate selection are clarified in Section 3.4.

Section 4 discusses the logical foundations of SEM and resolves a number of difficulties that were kept dormant in the past. These include operational definitions for structural equations, structural parameters, error terms, total and direct effects.

2 GRAPHS AND MODEL TESTING

In 1919, Wright developed his "method of path coefficients," which allows researchers to compute the magnitudes of cause-effect relationships from correlation measurements, as long as the path diagram represents correctly the causal processes underlying the data. Wright's method consists of writing a set of equations, one for each pair of variables (X_i, X_j), and equating the (standardized) correlation coefficient ρ_{ij} with a sum of products of path coefficients and residual correlations along the various paths connecting X_i and X_j. Whenever the resulting equations give a unique solution to some path coefficient p_{mn} that is independent of the (unobserved) residual correlations, that coefficient is said to be *identifiable*. If every set of correlation coefficients ρ_{ij} is compatible with some choice of path coefficients, the model is said to be *untestable* or *unfalsifiable* (also called *saturated, just identified*, and so on), because it is capable of perfectly fitting any data whatsoever.

Whereas Wright's method is partly graphical and partly algebraic, the theory of directed graphs permits us to analyze questions of testability and identifiability in purely graphical terms, prior to data collection, and it also enables us to extend these analyses from linear to nonlinear or nonparametric models. This section deals with issues of testability in linear and nonparametric models.

2.1 The Testable Implications of Structural Models

When we hypothesize a model of the data-generating process, that model often imposes restrictions on the statistics of the data collected. In observational studies, these restrictions provide the only view under which the hypothesized model can be tested or falsified. In many cases, such restrictions can be expressed in the form of vanishing partial correlations and, more significant, the restrictions are implied by the structure of the path diagram alone, independent of the numerical values of the parameters. Blalock [1962], having recognized the importance of vanishing partial correlations that are implied by path diagrams, worked out an exhaustive list of those correlations in all path diagrams involving four variables. He also expressed doubt that the list would ever be extended to path diagrams with five (or more) variables. Nonetheless, a method is now available for identifying vanishing partial correlations in path diagrams of any size or form. The method is based on a test called d-separation [Pearl, 1986, 1988], to be discussed next.

2.1.1 Preliminary notation

The graphs we discuss in this paper represent sets of structural equations of the form

$$x_i = f_i(pa_i, \epsilon_i) \qquad i = 1, \ldots, n \tag{1}$$

where pa_i (connoting *parents*) stand for the set of variables judged to be immediate causes of X_i, and ϵ_i represent errors due to omitted factors. Eq. (1) is a nonlinear, nonparametric generalization of the standard linear equations

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + \epsilon_i \qquad i = 1, \ldots, n \tag{2}$$

in which pa_i correspond to those variables on the r.h.s. of Eq. (2) that have non-zero coefficients. A set of equations in the form of Eq. (1) used to represent the data-generating process will be called a *causal model.*[7] The graph G obtained by drawing an arrow from every member of pa_i to X_i will be called a *causal diagram.* In addition to full arrows, a causal diagram should contain a bi-directed (i.e., double-arrowed) arc between any pair of variables whose corresponding errors are dependent (as in Fig. 3). A diagram may include directed cycles (e.g., $X \longrightarrow Y$, $Y \longrightarrow X$), representing mutual causation or feedback processes, but not self loops (e.g., $X \longrightarrow X$). An *edge* is either an arrow or a bi-directed arc, and two variables connected by an edge are called *adjacent.*

We make free use of the terminology of kinship (e.g., *parents, children, descendants, ancestors*) to denote the relationships in a graph. These kinship relations are defined along the full arrows in the diagram, including arrows that form directed cycles but ignoring bi-directed arcs. In Fig. 5(c), for example, Y has two parents (X and Z), three ancestors (X, Z, and W), and no children, while X has no parents (hence, no ancestors) and one child (Y).

[7]Causal models, structural equations, and error terms will be defined in terms of response to interventions in Section 4. Formal treatments of these notions are given in [Galles and Pearl, 1997, 1998].

Causal diagrams play the same role in nonlinear structural equation models as path diagrams play in linear structural equation models. Causal diagrams differ from path diagrams in that their pa_i are defined as nontrivial arguments of the function f_i, rather than variables obtaining non-zero coefficients, and their bi-directed arcs reflect dependency, rather than correlation. It is important to emphasize that causal diagrams (as well as traditional path diagrams) should be distinguished from the wide variety of graphical models in the statistical literature whose construction rests solely on properties of the joint distribution [Wermuth and Cox, 1996; Andersson et al., 1998; Lauritzen, 1997]. The missing links in those statistical models represent conditional independencies, while the missing links in causal diagrams represent absence of causal connections (see footnote 3 and Section 4) which may or may not imply conditional independencies in the distribution.

A causal model will be called *Markovian* if its graph contains no directed cycles and if its ϵ_i''s are mutually independent (i.e., no bi-directed arcs). A model is *semi-Markovian* if its graph is acyclic and if it contains dependent errors.

Markovian models, (the parallel term in the SEM literature is *recursive models*[8] [Bollen, 1989]), possess useful features, shared by both linear and nonlinear systems, that make their statistical implications transparent. One fundamental property of Markovian models is *parent screening*: given the state of its parents pa_i, each variable X_i is conditionally independent of all its nondescendants in the graph. This follows immediately from the independence of the errors ϵ_i and supports the intuition that once the direct causes of X_i are known, the probability of X_i is completely determined; no other event preceding X_i could modify this probability. As a result, the statistical parameters of Markovian models can be estimated by ordinary regression analysis.

An immediate consequence of this Markovian property is that the joint distribution of variables generated by Eq. (1) can be decomposed (using the chain rule of probability calculus) into the product

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid pa_i) \tag{3}$$

where pa_i are the values of the parents of X_i in the causal graph G. For example, the model illustrated in Fig. 1 induces the decomposition

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1)\ P(x_4|x_2, x_3)\ P(x_5|x_4) \tag{4}$$

This decomposition holds for any distribution of the error terms, regardless of the functional form of f_i; it depends only on the structural features of the generating model in Eq. (1) as captured by the graph G. The product decomposition, in turn, entails certain conditional independence relationships that hold regardless of the functional form of f_i and regardless of the error distribution. Such independencies are said to be *entailed* by the graph and can be read from the graph using a criterion called *d-separation* (the d denotes *directional*).

2.1.2 The d-separation criterion

Consider three disjoint sets of variables, X, Y, and Z, which are represented as nodes in a directed acyclic graph (DAG) G. To test whether X is independent of Y given Z in any

[8] The term *recursive* is ambiguous; some authors exclude correlated errors, but others do not.

Markovian model represented by G, we need to test whether the nodes corresponding to variables Z "block" all paths from nodes in X to nodes in Y. By *path* we mean a sequence of consecutive edges (of any directionality) in the graph, and "blocking" is to be interpreted as stopping the flow of information (or the correlation) between the variables that are connected by such paths.

Definition 1 (d-separation) *A path p is said to be* d-separated (or blocked) *by a set of nodes Z iff*

1. *p contains a chain $i \longrightarrow m \longrightarrow j$ or a fork $i \longleftarrow m \longrightarrow j$ such that the middle node m is in Z, or*
2. *p contains an inverted fork (or* collider*) $i \longrightarrow m \longleftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.*

A set Z is said to d-separate X from Y iff Z blocks every path from a node in X to a node in Y.

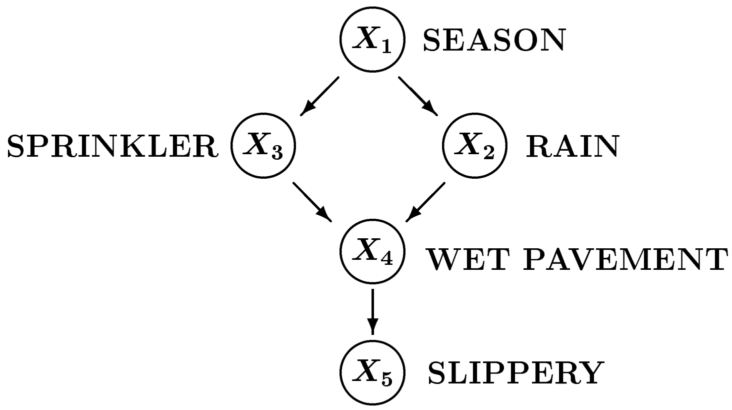

Figure 1: Graph illustrating causal relationships among five variables.

The intuition behind d-separation is simple. In causal chains $i \longrightarrow m \longrightarrow k$ and causal forks $i \longleftarrow m \longrightarrow j$, the two extreme variables are marginally dependent but become independent of each other (i.e., blocked) once we condition on the middle variable. Figuratively, conditioning on m appears to "block" the flow of information along the path, since learning about i has no effect on the probability of j, given m. Inverted forks $i \longrightarrow m \longleftarrow j$, representing two causes having a common effect, act the opposite way; if the two extreme variables are (marginally) independent, they will become dependent (i.e., connected through unblocked path) once we condition on the middle variable (i.e., the common effect) or any of its descendants. This can be confirmed in the context of Fig. 1. Once we know the season, X_3 and X_2 are independent (assuming that sprinklers are set in advance, according to the season) but finding that the pavement is wet or slippery renders X_2 and X_3 dependent, because refuting one of these explanations increases the probability of the other.

In Fig. 1, $X = \{X_2\}$ and $Y = \{X_3\}$ are d-separated by $Z = \{X_1\}$, because both paths connecting X_2 and X_3 are blocked by Z. The path $X_2 \leftarrow X_1 \rightarrow X_3$ is blocked because it is a fork in which the middle node, X_1, is in Z, while the path $X_2 \rightarrow X_4 \leftarrow X_3$ is blocked because it is an inverted fork in which the middle node, X_4, and all its descendants are outside Z. However, X and Y are not d-separated by the set $Z' = \{X_1, X_5\}$: the path $X_2 \rightarrow X_4 \leftarrow X_3$ (an inverted fork) is not blocked by Z', since X_5, a descendant of the middle node X_4, is in Z'. Metaphorically, learning the value of the consequence X_5 renders its causes X_2 and X_3 dependent, as if a pathway were opened along the arrows converging at X_4.

Readers might find it a bit odd that conditioning on a node not lying on a blocked path may unblock the path. However, this corresponds to a general rule about causal relationships: observations on a common consequence of uncorrelated causes tend to render those causes correlated. This rule is known as *Berkson's paradox* in the statistical literature [Berkson, 1946] and as the *explaining away effect* in artificial intelligence [Kim and Pearl, 1983]. For example, if the admission criteria to a certain graduate school call for either high grades as an undergraduate or special musical talents, then these two attributes will be found to be negatively correlated in the student population of that school, even if these attributes are uncorrelated in the population at large. Indeed, students with low grades are likely to be exceptionally gifted in music, which explains their admission to graduate school.

Algebraically, the partial correlations associated with $i \rightarrow m \leftarrow j$ are governed by the equation $\rho_{ij \cdot m} = (\rho_{ij} - \rho_{im}\rho_{jm})/(1 - \rho_{im}^2)^{\frac{1}{2}}(1 - \rho_{jm}^2)^{\frac{1}{2}}$ which renders $\rho_{ij \cdot m} \neq 0$ when $\rho_{ij} = 0$. The same applies to the partial correlation $\rho_{ij \cdot m'}$ where m' is any descendant of m.

Theorem 1 [Verma and Pearl, 1988; Geiger et al., 1990] *If sets X and Y are d-separated by Z in a DAG G then X is independent of Y conditional on Z in every Markovian model structured according to G. Conversely, if X and Y are* not *d-separated by Z in a DAG G, then X and Y are dependent conditional on Z in* almost all *Markovian models structured according to G.*

Because conditional independence implies zero partial correlation, Theorem 1 translates into a graphical test for identifying those partial correlations that must vanish in the model.

Corollary 1 *In any Markovian model structured according to a DAG G, the partial correlation $\rho_{XY \cdot Z}$ vanishes whenever the nodes corresponding to the variables in Z d-separate node X from node Y in G, regardless of the model's parameters. Moreover, no other partial correlation would vanish for all the model's parameters.*

Unrestricted semi-Markovian models can always be emulated by Markovian models that include latent variables, with the latter accounting for all dependencies among error terms. Consequently, the d-separation criterion remains valid in such models if we interpret bi-directed arcs as emanating from latent common parents. This is not always possible in linear semi-Markovian models if each latent variable is restricted to influence at most two observed variables [Spirtes et al., 1996]. However, it has been shown that the d-separation criterion remains valid in such restricted systems [Spirtes et al., 1996] and, moreover, that the validity is preserved when the network contains cycles [Koster, 1998; Spirtes et al., 1998]. These results are summarized in the next theorem.

Theorem 2 *For any linear model structured according to diagram D, which may include cycles and bi-directed arcs, the partial correlation $\rho_{XY \cdot Z}$ vanishes if the nodes corresponding to the set of variables Z d-separate node X from node Y in D, where each bi-directed arc $i \dashleftarrow\dashrightarrow j$ is interpreted as a latent common parent $i \longleftarrow L \longrightarrow j$.*

For linear structural equation models (see Eq. (2)), Theorem 2 implies that those (and only those) partial correlations identified by the d-separation test are guaranteed to vanish independent of the model parameters α_{ik} and independent of the error variances. This suggests a simple and direct method for testing models: rather than going through the standard exercise of finding a maximum likelihood estimate for the model's parameters and scoring those estimates for fit to the data, we can directly test for each zero partial correlation implied by the free model. The advantages of using such tests were noted by Shipley [1997], who also devised implementations of these tests.

The question arises however whether it is feasible to test for the vast number of vanishing partial correlations entailed by a given model. Fortunately, these partial correlations are not independent of each other, but can be derived from a relatively small number of partial correlations that constitutes a *basis* for the entire set [Pearl and Verma, 1987].

Definition 2 (basis) *Let S be a set of partial correlations. A basis B for S is a set of zero partial correlations that (1) implies (using the laws of probability) the vanishing of every element of S, and (2) no proper subset of B sustains such implication.*

An obvious choice of a basis for the zero partial correlations entailed by a DAG D is the set of equalities $B = \{\rho_{ij \cdot pa_i} = 0 | i > j\}$, where i ranges over all nodes in D, and j ranges over all predecessors of i in any order that agrees with the arrows of D. This set of equalities reflects in fact the "parent screening" property of Markovian models, which is the source of all the probabilistic information encoded in a DAG . Testing for these equalities is sufficient therefore for testing all the statistical claims of a linear Markovian model. Moreover, when the parent sets pa_i are large, it may be possible to select a more economical basis, as shown in the next theorem.[9]

Theorem 3 *(Graphical basis) Let (i, j) be a pair of nonadjacent nodes in a DAG D, and Z_{ij} any set of nodes that are closer to i than j is to i, and such that Z_{ij} d-separates i from j. The set of zero partial correlations $B = \{\rho_{ij \cdot Z_{ij}} = 0 | i > j\}$, consisting of one element per nonadjacent pair, constitutes a basis for the set of all vanishing partial correlations entailed by D.*

Theorem 3 states that the set of zero partial correlations corresponding to *any* separation between nonadjacent nodes in the diagram encapsulates the entire statistical information conveyed by a linear Markovian model. A formal proof of Theorem 3 is given in [Pearl and Meshkat, 1998].

[9]The possibility that linear models may possess more economical bases came to my awareness during a conversation with Rod McDonald.

Examining Fig. 2, we see that each of following two sets forms a basis for the model in the figure:

$$\begin{aligned} B_1 &= \{\rho_{32\cdot 1} = 0, \rho_{41\cdot 3} = 0, \rho_{42\cdot 3} = 0, \rho_{51\cdot 43} = 0, \rho_{52\cdot 43} = 0\} \\ B_2 &= \{\rho_{32\cdot 1} = 0, \rho_{41\cdot 3} = 0, \rho_{42\cdot 1} = 0, \rho_{51\cdot 3} = 0, \rho_{52\cdot 1} = 0\} \end{aligned} \tag{5}$$

The basis B_1 employs the parent set pa_i for separating i from j, $i > j$. Basis B_2, on the other hand, employs smaller separating sets, thus leading to tests involving fewer regressors. Note that each member of a basis corresponds to a missing arrow in the DAG; therefore, the number of tests required to validate a DAG is equal to the number of missing arrows it contains. The sparser the graph, the more it constrains the covariance matrix and more tests are required to verify those constraints.

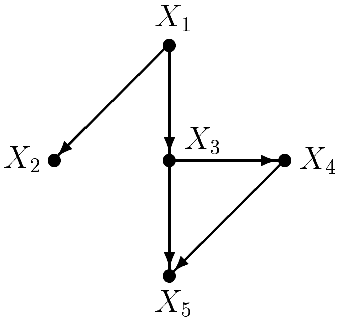

Figure 2: Model testable with two regressors for each missing link (Eq. (5))

2.2 Testing the Testable

In linear structural equation models, the hypothesized causal relationships between variables can be expressed in the form of a directed graph annotated with coefficients, some fixed a priori (usually to zero) and some free to vary. The conventional method for testing such a model against the data involves two stages. First, the free parameters are estimated by iteratively maximizing a fitness measure such as the maximum likelihood function. Second, the covariance matrix implied by the estimated parameters is compared to the sample covariances and a statistical test is applied to decide whether the latter could originate from the former [Bollen, 1989; Chou and Bentler, 1995].

There are two major weaknesses to this approach:

1. If some parameters are not identifiable, the first phase may fail to reach stable estimates for the parameters and the investigator must simply abandon the test.

2. If the model fails to pass the data-fitness test, the investigator receives very little guidance about which modeling assumptions are wrong.

For example, Fig. 3 shows a path model in which the parameter α is not identifiable if $cov(\epsilon_1, \epsilon_2)$ is assumed unknown, which means that the maximum likelihood method may fail to find a suitable estimate for α, thus precluding the second phase of the test. Still, this model is no less testable than the one in which $cov(\epsilon_1, \epsilon_2) = 0$, α is identifiable, and the test

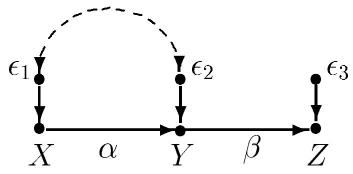

Figure 3: A testable model containing unidentified parameter (α)

can proceed. These models impose the same restrictions on the covariance matrix, namely, that the partial correlation $\rho_{XZ \cdot Y}$ should vanish (i.e., $\rho_{XZ} = \rho_{XY}\rho_{YZ}$), yet the model with free $cov(\epsilon_1, \epsilon_2)$, by virtue of α being nonidentifiable, cannot be tested for this restriction.

Fig. 4 illustrates the weakness associated with model diagnosis. Suppose the true data-generating model has a direct causal connection between X and W, as shown in Fig. 4(a), while the hypothesized model (Fig. 4(b)) has no such connection. Statistically, the two models differ in the term $\rho_{XW \cdot Z}$, which should vanish according to Fig. 4(b) and is left free according to Fig. 4(a). Once the nature of the discrepancy is clear, the investigator must

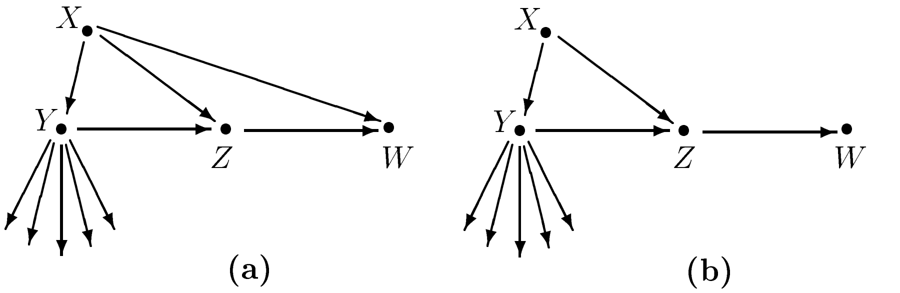

Figure 4: Models differing in one local test, $\rho_{XW \cdot Z} = 0$

decide whether substantive knowledge justifies alteration of the model, namely, adding either a link or a curved arc between X and W. However, because the effect of the discrepancy will be spread over several covariance terms, global fitness tests will not be able to isolate the discrepancy easily. Even multiple fitness tests on various local modifications of the model (such tests are provided by LISREL) may not help much, because the results may be skewed by other discrepancies in different parts of the model, such as the subgraph rooted at Y. Thus, testing for global fitness is often of only minor use in model debugging.

Local fitness testing is an attractive alternative to global fitness testing. This involves listing the restrictions implied by the model and testing them one by one. Local testing may help isolate the discrepancy and can be performed more reliably than testing the overall model as one unit. A restriction such as $\rho_{XW \cdot Z} = 0$, for example, can be tested locally without measuring Y or any of its descendants, thus keeping errors associated with those measurements from interfering with the test for $\rho_{XW \cdot Z} = 0$, which is the real source of the lack of fit. More generally, typical SEM models are often close to being "saturated," claiming but a few restrictions, in the form of a few edges missing from large, otherwise unrestrictive diagrams. Local and direct tests for those restrictions are more reliable than global tests, as they involve fewer degrees of freedom and are not contaminated with irrelevant measurement errors. The missing edges approach described in section 2.1 provides a systematic way of detecting and enumerating the local tests needed for testing a given model.

2.3 Model Equivalence

An important criterion for determining whether two given causal models are observationally equivalent follows from the d-separation test.

Definition 3 (observational equivalence) *Two structural equation models are said to be* observationally equivalent *if every probability distribution that is generated by one of the models can also be generated by the other.*

Theorem 4 [Verma and Pearl, 1990] *Two Markovian models are observationally equivalent iff they entail the same sets of conditional independencies. Moreover, two such models are observationally equivalent iff their corresponding graphs have the same sets of edges and the same sets of v-structures (two converging arrows whose tails are not connected by an arrow).*

In standard SEM, models are assumed linear and data are characterized by covariance matrices. Thus, two such models are observationally indistinguishable if they are *covariance equivalent*, that is, if every covariance matrix generated by one model (through some choice of parameters) can also be generated by the other. It can be easily verified that Theorem 4 extends to covariance equivalence.

Theorem 5 *Two Markovian linear-normal models are covariance equivalent iff they entail the same sets of zero partial correlations. Moreover, two such models are covariance equivalent iff their corresponding graphs have the same sets of edges and the same sets of v-structures.*

In Theorems 4 and 5, the first part defines the testable implications of any Markovian structural equation model. These theorems state that, in nonmanipulative studies, Markovian structural equation models cannot be tested for any feature other than those zero partial correlations that the d-separation test reveals. They provide as well a simple test for equivalence which requires, instead of the checking of all d-separation conditions, merely a comparison of corresponding edges and their directionalities.

For example, reversing the direction of the arrow between X_1 and X_2 in Fig. 1 does not introduce any new v-structure. Therefore, this reversal yields an observationally equivalent network, and the directionality of the link $X_1 \longrightarrow X_2$ cannot be determined from statistical information. The arrows $X_2 \longrightarrow X_4$ and $X_4 \longrightarrow X_5$ are of a different nature, however; their directionality cannot be reversed without creating a new v-structure. Thus, we see that some arrows retain their directionality in all models equivalent to a given model and, hence, that this directionality is testable whenever the equivalence class (of models) is testable. Algorithms for automatically identifying such arrows in the graph have been devised by Chickering [1995], Meek [1995], and Andersson et al. [1998]. We further see that some kinds of statistical data (such as those generated by the model in Fig. 1), unaccompanied by temporal information, can reveal the directionality of some arrows and, hence, the directionality of the causal relationships among the corresponding variables. This feature is used in various discovery algorithms that elicit causal relationships from complex patterns of statistical associations (e.g., [Pearl and Verma, 1991; Spirtes et al., 1993]), but discussion of such algorithms lies beyond the scope of this paper.

In semi-Markovian models (DAGs with correlated errors), the d-separation criterion is still valid for testing independencies (see Theorem 2) but independence equivalence no longer implies observational equivalence.[10] Two models that entail the same set of zero partial correlations among the observed variables may yet impose different inequality constraints on the covariance matrix. Nevertheless, Theorems 2 and 4 still provide necessary conditions for testing equivalence.

2.3.1 Generating equivalent models

By permitting arrows to be reversed as long as no v-structures are destroyed or created, we can use Theorems 4 and 5 to generate equivalent alternatives to any Markovian model. Meek [1995] and Chickering [1995] have shown that $X \longrightarrow Y$ can be replaced by $X \longleftarrow Y$ iff all parents of X are also parents of Y, and, moreover, that for any two equivalent models, there is always some sequence of such edge reversals that takes one model into the other. This simple rule for edge reversal coincides with those proposed by Stelzl [1986] and Lee and Hershberger [1990].

In semi-Markovian models, the rules for generating equivalent models are more complicated. Nevertheless, Theorems 4 and 5 yield convenient graphical principles for testing the correctness of edge-replacement rules.

The basic principle is that if we regard each bi-directed arc $X \dashleftarrow\dashrightarrow Y$ as representing a latent common cause $X \longleftarrow L \longrightarrow Y$, then the "if" part of Theorem 4 remains valid, that is, any edge-replacement transformation that does not destroy or create a v-structure is allowed. Thus, for example, an edge $X \longrightarrow Y$ can be replaced by a bi-directed arc $X \dashleftarrow\dashrightarrow Y$ whenever X and Y have no other parents, latent or observed. Likewise, an edge $X \longrightarrow Y$ can be replaced by a bi-directed arc $X \dashleftarrow\dashrightarrow Y$ whenever (1) X and Y have no latent parents and (2) every parent of X or Y is a parent of both. Such replacements do not introduce new v-structures. Since v-structures may now involve latent variables, however, we can tolerate the creation or destruction of some v-structures as long as this does not effect partial correlations among the observed variables. Fig. 5(a) demonstrates that the creation of certain v-structures can be tolerated. If we reverse the arrow $X \longrightarrow Y$ we create two converging arrows $Z \longrightarrow X \longleftarrow Y$ whose tails are connected, not directly, but through a latent common cause. This is tolerated, because, although the new convergence at X blocks the path (Z, X, Y), the connection between Z and Y (through the arc $Z \dashleftarrow\dashrightarrow Y$) remains unblocked and, in fact, cannot be blocked by any set of observed variables.

We can carry this principle further by generalizing the concept of v-structure. Whereas in Markovian models, a v-structure is defined as two converging arrows whose tails are not connected by a link, we now define v-structure as any two converging arrowheads whose tails are *separable*. By separable, we mean that there exists a conditioning set S capable of d-separating the two tails. Clearly, the two tails will not be separable if they are connected by an arrow or by a bi-directed arc. But a pair of nodes in a semi-Markovian model can be inseparable even when not connected by an edge [Verma and Pearl, 1990]. With this generalization in mind, we can state necessary conditions for edge replacement:

[10] Verma and Pearl [1990] present an example using a nonparametric model, and Richardson has devised an example using linear models with correlated errors [Spirtes and Richardson, 1996].

Rule 1: An arrow $X \longrightarrow Y$ is interchangeable with $X \dashleftarrow\dashrightarrow Y$ only if every neighbor or parent of X is inseparable from Y. (By *neighbor* we mean a node connected (to X) through a bi-directed arc.)

Rule 2: An arrow $X \longrightarrow Y$ can be reversed into $X \longleftarrow Y$ only if, before reversal, (i) every neighbor or parent of Y (excluding X) is inseparable from X and (ii) every neighbor or parent of X is inseparable from Y.

For example, consider the model $Z \dashleftarrow\dashrightarrow X \longrightarrow Y$. The arrow $X \longrightarrow Y$ cannot be replaced with a bi-directed arc $X \dashleftarrow\dashrightarrow Y$ because Z (a neighbor of X) is separable from Y by the set $S = \{X\}$. Indeed, the new v-structure created at X would render X and Y marginally independent, contrary to the original model.

As another example, consider the graph in Fig. 5(a). Here, it is legitimate to replace

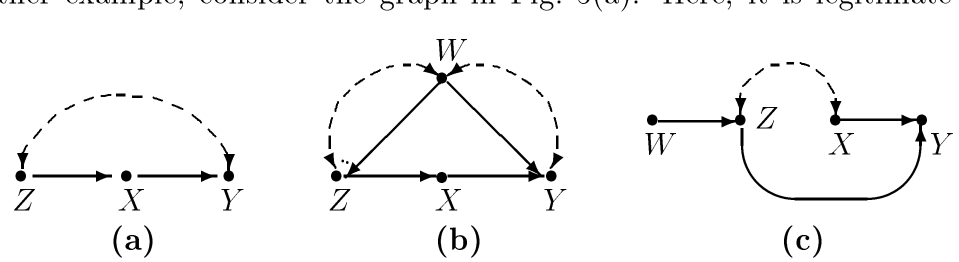

Figure 5: Models permitting ((a) and (b)) and forbidding (c) the reversal of $X \rightarrow Y$

$X \longrightarrow Y$ with $X \dashleftarrow\dashrightarrow Y$ or with a reversed arrow $X \longleftarrow Y$ because X has no neighbors and Z, the only parent of X, is inseparable from Y. The same considerations apply to Fig. 5(b); variables Z and Y, though nonadjacent, are inseparable, because the paths going from Z to Y through W cannot be blocked.

A more complicated example, one that demonstrates that the rules above are not sufficient to ensure the legitimacy of a transformation, is shown in Fig. 5(c). Here, it appears that replacing $X \longrightarrow Y$ with $X \dashleftarrow\dashrightarrow Y$ would be legitimate because the (latent) v-structure at X is shunted by the arrow $Z \longrightarrow Y$. However, the original model shows the path from W to Y to be d-connected given Z, while the post-replacement model shows the same path d-separated given Z. Consequently, the partial correlation $\rho_{WY \cdot Z}$ vanishes in the post-replacement model but not in the pre-replacement model. A similar disparity also occurs relative to the partial correlation $\rho_{WY \cdot ZX}$. The original model shows that the path from W to Y is blocked, given $\{Z, X\}$, while the post-replacement model shows that path d-connected, given $\{Z, X\}$. Consequently, the partial correlation $\rho_{WY \cdot ZX}$ vanishes in the pre-replacement model but is unconstrained in the post-replacement model.[11] Evidently, it is not enough to impose rules on the parents and neighbors of X; remote ancestors (e.g., W) should be considered too.

These rules are just a few of the implications of the d-separation criterion when applied to semi-Markovian models. A necessary and sufficient criterion for testing the d-separation equivalence of two semi-Markovian models has been devised by Spirtes and Verma [1992]. Spirtes and Richardson [1996] have extended that criterion to include models with feedback cycles. We should keep in mind, though, that, because two semi-Markovian models can

[11] This example was brought to my attention by Jin Tian, and a similar one, by two anonymous reviewers.

be zero-partial-correlation equivalent and yet not covariance equivalent, criteria based on d-separation can provide merely the necessary conditions for model equivalence.

2.3.2 The significance of equivalent models

Theorem 4 is methodologically significant because it clarifies what it means to claim that structural models are "testable" [Bollen, 1989, p. 78].[12] It asserts that we never test *a* model but, rather a whole class of observationally equivalent models from which the hypothesized model cannot be distinguished by any statistical means. It asserts as well that this equivalence class can be constructed by inspection, from the graph, which thus provides the investigator with a vivid representation of competing alternatives for consideration. Graphs representing all models in a given equivalence class have been given by Verma and Pearl [1990], Spirtes et al. [1993], and Andersson et al. [1998]. Richardson [1996] discusses the representation of equivalence classes of models with cycles.

While it is true that (over-identified) structural equation models have testable implications, those implications are but a small part of what the model represents, namely, a set of claims, assumptions, and implications. Failure to distinguish among causal assumptions, statistical implications, and policy claims has been one of the main reasons for the suspicion and confusion surrounding quantitative methods in the social sciences [Freedman, 1987, p. 112; Goldberger, 1992; Wermuth, 1992]. However, because they make the distinctions among these components vivid and crisp, graphical methods promise to make SEM more acceptable to researchers from a wide variety of disciplines.

By and large, the SEM literature has ignored the explicit analysis of equivalent models. Breckler [1990], for example, found that out of 72 articles in the areas of social and personality psychology only one acknowledged the existence of an equivalent model. The general attitude has been that the combination of data fitness and model over-identification is sufficient to confirm the hypothesized model. Recently, however, the existence of multiple equivalent models seems to have jangled the nerves of some SEM researchers. MacCallum et al. [1993, p. 198] conclude that "the phenomenon of equivalent models represents a serious problem for empirical researchers using CSM" and "a threat to the validity of interpretation of CSM results." Breckler [1990, p. 262] reckons that "if one model is supported, so too are all of its equivalent models" and, consequently, ventures that "the term *causal modeling* is a misnomer."

Such extremes are not justifiable. The existence of equivalent models is logically inevitable if we accept the fact that causal relations cannot be inferred from statistical data alone; as Wright [1921] stated, "prior knowledge of the causal relations is assumed as prerequisite" in SEM. But this does not make SEM useless as a tool for causal modeling. The move from the qualitative causal premises represented by the structure of a path diagram (see footnote 3) to the quantitative causal conclusions advertised by the coefficients in the diagram is neither useless nor trivial. Consider, for example, the model depicted in Fig. 6, which Bagozzi and Burnkrant [1979] use to illustrate problems associated with equivalent models. Although this model is saturated (i.e., just identified) and although it has (at least)

[12] In response to an allegation that "path analysis does not derive the causal theory from the data, or test any major part of it against the data" [Freedman, 1987, p. 112], Bollen [1989, p. 78] states, "we can test and reject structural models.... Thus the assertion that these models cannot be falsified has little basis."

27 semi-Markovian equivalent models, finding that the influence of AFFECT on BEHAVIOR is almost three times stronger (on a standardized scale) than the influence of COGNITION on BEHAVIOR is still very illuminating — it tells us about the relative effectiveness of different behavior-modification policies if some are known to influence AFFECT and others COGNITION. The significance of this quantitative analysis on policy analysis may be more dramatic when a path coefficient turns negative while the corresponding correlation coefficient measures positive. Learning that such a reversal is logically implied by the qualitative causal premises embedded in the diagram may have profound impact on policy decisions.

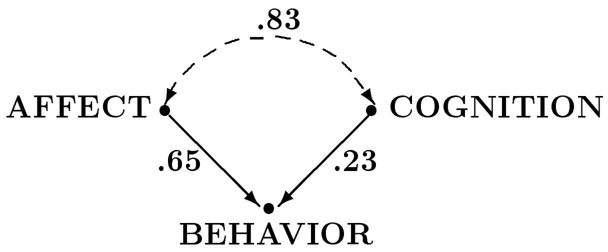

Figure 6: Untestable model displaying quantitative causal information derived

In summary, social scientists need not abandon SEM altogether; they need only abandon the notion that SEM is a method of *testing* causal models. SEM is a method of testing a tiny fraction of the premises that make up a causal model and, in cases where that fraction is found to be compatible with the data, the method elucidates the necessary quantitative consequences of both the premises and the data. It follows, then, that users of SEM should concentrate on examining the implicit theoretical premises that enter into a model. As we will see in Section 4, graphical methods make these premises vivid and precise.

3 GRAPHS AND IDENTIFIABILITY

3.1 Parameter Identification in Linear Models

Consider a directed edge $X \longrightarrow Y$ embedded in a path diagram G, and let α stand for the path coefficient associated with that edge. It is well known that the regression coefficient $r_{YX} = \rho_{XY}\sigma_Y/\sigma_X$ can be decomposed into the sum

$$r_{YX} = \alpha + I_{YX}$$

where I_{YX} is not a function of α, since it is computed (e.g., using Wright's rules) from other paths connecting X and Y excluding the edge $X \longrightarrow Y$. (Such paths traverse both uni-directed and bi-directed arcs.) Thus, if we remove the edge $X \longrightarrow Y$ from the path diagram and find that the resulting subgraph entails zero correlation between X and Y, then we know that $I_{YX} = 0$ and $\alpha = r_{YX}$; hence, α is identified. Such entailment can be established graphically by testing whether X is d-separated from Y (by the empty set $Z = \{\emptyset\}$) in the subgraph. Fig. 7 illustrates this simple test for identification: all paths between X and Y in the subgraph G_α are blocked by converging arrows, and α can immediately be equated with r_{YX}.

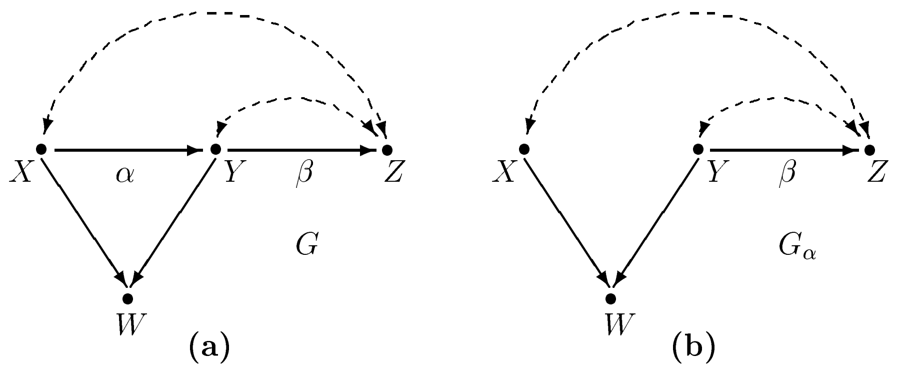

Figure 7: Testing whether structural parameter α can be equated with regression coefficient r_{YX}

We can extend this basic idea to cases where I_{YX} is not zero but can be made zero by adjusting for a set of variables $Z = \{Z_1, Z_2, \ldots, Z_k\}$ that lie on various d-connected paths between X and Y. Consider the partial regression coefficient $r_{YX \cdot Z} = \rho_{YX \cdot Z}\sigma_{Y \cdot Z}/\sigma_{X \cdot Z}$, which represents the residual correlation between Y and X after Z is "partialled out." If Z contains no descendant of Y, then again we can write[13]

$$r_{YX \cdot Z} = \alpha + I_{YX \cdot Z}$$

where $I_{YX \cdot Z}$ represents the partial correlation between X and Y resulting from setting α to zero, that is, the partial correlation in a model whose graph, G_α, lacks the edge $X \longrightarrow Y$ but is otherwise identical to G. If Z d-separates X from Y in G_α, then $I_{YX \cdot Z}$ would indeed be zero in such a model, and we can conclude that in our original model, α is identified and is equal to $\alpha = r_{YX \cdot Z}$. Moreover, since $r_{YX \cdot Z}$ is given by the coefficient of x in the regression of Y on X and Z, α can be estimated using the regression

$$y = \alpha x + \beta_1 z_1 + \ldots + \beta_k z_k + \epsilon$$

This result provides a simple graphical answer to the questions, alluded to in Section 1.3, of what constitutes an adequate set of regressors and when a regression coefficient provides a consistent estimate of a path coefficient. The answers are summarized in the following theorem.[14]

Theorem 6 (single-link criterion) *Let G be any path diagram in which α is the path coefficient associated with link $X \longrightarrow Y$, and let G_α denote the diagram that results when $X \longrightarrow Y$ is deleted from G. The coefficient α is identifiable, if there exists a set of variables Z such that Z contains no descendant of Y, and Z d-separates X from Y in G_α. If Z satisfies these two conditions, then α is equal to the regression coefficient $r_{YX \cdot Z}$. Conversely, if Z does not satisfy these conditions, then $r_{YX \cdot Z}$ is not a consistent estimand of α, except in rare instances of measure zero.*

[13] This can be seen when the relation between Y and its parents, $Y = \alpha x + \sum_i \beta_i w_i + \epsilon$ is substituted into the expression for $r_{YX \cdot Z}$, which yields α plus an expression $I_{YX \cdot Z}$ involving partial correlations among the variables $\{X, W_1, \ldots, W_k, Z, \epsilon\}$. Since Y is assumed not to be an ancestor of any of these variables, their joint density is unaffected by the equation for Y; hence, $I_{YX \cdot Z}$ is independent of α.

[14] This result is also presented in [Spirtes et al., 1998].

The use of Theorem 6 can be illustrated as follows. Consider the graphs G and G_α in Fig. 8. The only path connecting X and Y in G_α is the one traversing Z, and since that path

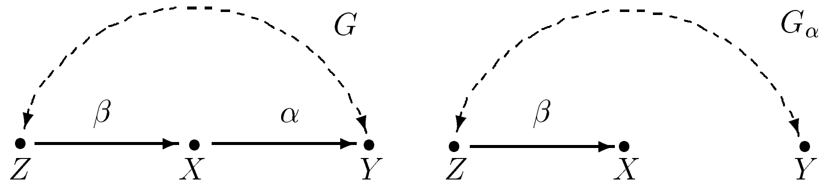

Figure 8: Illustrating the Identification of α (Theorem 6)

is d-separated (blocked) by Z, α is identifiable and is given by $\alpha = r_{YX \cdot Z}$. The coefficient β is identifiable, of course, since Z is d-separated from X in G_β (by the empty set $\{\emptyset\}$) and thus $\beta = r_{XZ}$.

We now extend the use of d-separation to facilitate the identification of total, rather than direct, effects. Consider the graph G in Fig. 9. If we form the graph G_α (by removing the

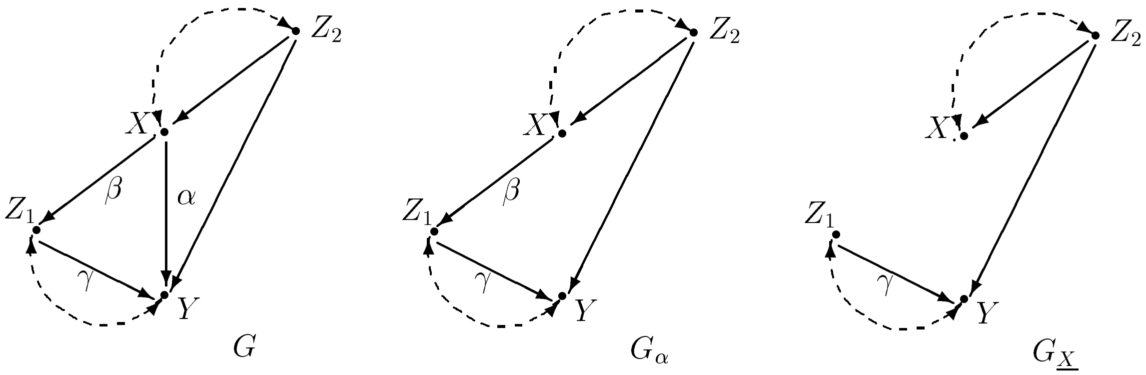

Figure 9: Graphical identification of the total effect of X on Y, $\alpha + \beta\gamma = r_{YX \cdot Z_2}$

link $X \longrightarrow Y$), we observe that there is no set Z of nodes that d-separates all paths from X to Y. If Z contains Z_1, then the path $X \longrightarrow Z_1 \leftarrow\!-\,-\!\rightarrow Y$ will be unblocked through the converging arrows at Z_1. If Z does not contain Z_1, the path $X \longrightarrow Z_1 \longrightarrow Y$ is unblocked. Thus we conclude that α cannot be identified using our previous method. However, suppose we are interested in the total effect of X on Y, given by $\alpha + \beta\gamma$. For this sum to be identified by r_{YX}, there should be no contribution to r_{YX} from paths other than those leading from X to Y. However, we see that two such paths, called *confounding* or *back-door* paths, exist in the graph, namely, $X \longleftarrow Z_2 \longrightarrow Y$ and $X \leftarrow\!-\,-\!\rightarrow Z_2 \longrightarrow Y$. Fortunately, these paths are blocked by Z_2, and we conclude that adjusting for Z_2 would render $\alpha + \beta\gamma$ identifiable and given by

$$\alpha + \beta\gamma = r_{YX \cdot Z_2}$$

This line of reasoning leads to a general test for the identifiability of total effects, called the *back-door* criterion [Pearl, 1993a; Pearl, 1995]:

Theorem 7 (back-door criterion) *For any two variables X and Y in a causal diagram G, the total effect of X on Y is identifiable if there exists a set of measurements Z such that*

1. *no member of Z is a descendant of X, and*

2. *Z d-separates X from Y in the subgraph $G_{\underline{X}}$ formed by deleting from G all arrows emanating from X.*

Moreover, if the two conditions are satisfied, then the total effect of X on Y is given by $r_{YX \cdot Z}$.

The two conditions of Theorem 7, as we will see in the next subsection, are also valid in nonlinear non-Gaussian models, as well as in models with discrete variables. It is for this reason that the back-door criterion can serve as a general test for covariate selection, as described in Section 1.3. The reason for deleting the arrows in step 2 is to ensure that only confounding (i.e., back-door) paths participate in the d-separation test. The test ensures that, after adjustment for Z, X and Y are not associated through confounding paths, which means that the partial correlation $r_{YX \cdot Z}$ is equal to the total effect. In fact, we can view Theorems 5 and 6 as special cases of a more general scheme: to identify any *partial effect*, as defined by a select bundle of causal paths from X to Y, we ought to find a set Z of measured variables that block all non-selected paths between X and Y. The partial effect will then equal the regression coefficient $r_{XY \cdot Z}$.

Figure 9 demonstrates that some total effects can be determined directly from the graphs, without having to identify their individual components. Standard SEM methods [Bollen, 1989; Chou and Bentler, 1995], which focus on the identification and estimation of individual parameters, may miss the identification and estimation of effects such as the one in Fig. 9, which can be estimated reliably even though some of the constituents remain unidentified.

Some total effects cannot be determined directly, as a unit, but require the determination of each component separately. In Fig. 8, for example, the effect of Z on $Y (= \alpha\beta)$ does not meet the back-door criterion, yet this effect can be determined from its constituents α and β, which meet the back-door criterion individually and evaluate to

$$\beta = r_{XZ} \qquad \alpha = r_{YX \cdot Z}$$

There is still a third kind of causal parameter which cannot be determined directly or through its constituents, but requires the evaluation of a broader causal effect of which it is a part. The structure shown in Fig. 10 represents an example of this case. The parameter

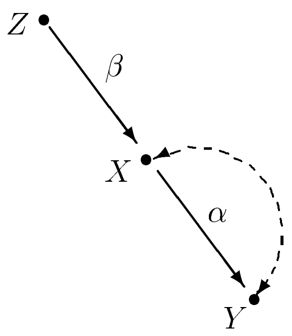

Figure 10: Graphical identification of α using instrumental variable Z

α cannot be identified directly, yet it can be determined from $\alpha\beta$ and β, which represent the effect of Z on Y and that of Z on X, respectively. These two effects can be identified directly, since there are no back-door paths from Z to either Y or X, giving $\alpha\beta = r_{YZ}$ and $\beta = r_{XZ}$. Thus,

$$\alpha = r_{YZ}/r_{XZ}$$

which is familiar to us as the instrumental-variable formula [Bowden and Turkington, 1984].

The example shown in Fig. 11 combines all three methods considered thus far. The total

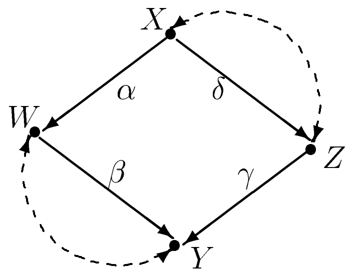

Figure 11: Graphical Identification of α, β and γ

effect of X on Y is given by $\alpha\beta + \gamma\delta$, which is not identifiable because it does not meet the back-door criterion and is not part of another identifiable structure. However, suppose we wish to estimate β. By conditioning on Z, we block all paths going through Z and obtain $\alpha\beta = r_{YX\cdot Z}$, which is the effect of X on Y mediated by W. Because there are no back-door paths from X to W, α itself evaluates directly to $\alpha = r_{WX}$. We therefore obtain

$$\beta = r_{YX\cdot Z}/r_{WX}$$

In contrast, γ can be evaluated directly by conditioning on X (thus blocking the back-door path from Z to Y through X), which gives

$$\gamma = r_{YZ\cdot X}$$

The methods we have been using suggest a systematic procedure for recognizing identifiable coefficients in a graph.

1. Start by searching for identifiable causal effects among pairs of variables in the graph, using the back-door criterion and Theorem 6. These can be either direct effects, total effects, or partial effects, that is, effects mediated by specific sets of variables.

2. For any such identified effect, collect the path coefficients involved and put them in a bucket.

3. Begin labeling the coefficients in the buckets according to the following procedure:
   If a bucket is a singleton, label its coefficient I (denoting *identifiable*).
   If a bucket is not a singleton but contains only a single unlabeled element, label that element I.

4. Repeat this process until no new labeling is possible.

5. List all labeled coefficients; these are identifiable.

The process described above is not complete, because our insistence on labeling coefficients one at a time may cause us to miss certain opportunities. This is shown in Fig. 12. Starting with the pairs $(X, Z), (X, W), (X', Z)$, and (X', W), we discover that α, γ, α', and

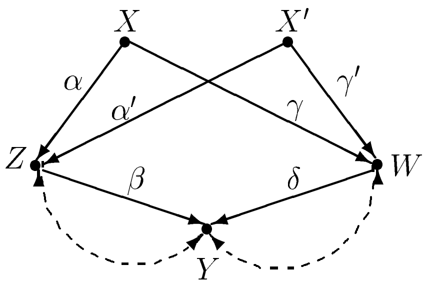

Figure 12: Identifying β and δ using two instrumental variables

γ' are identifiable. Going to (X, Y), we find that $\alpha\beta + \delta\gamma$ is identifiable, and, likewise, from (X', Y), that $\alpha'\beta + \gamma'\delta$ is identifiable. This does not enable us to label β or δ yet, but we can solve two equations for the unknowns β and δ, as long as the determinant $\begin{vmatrix} \alpha & \gamma \\ \alpha' & \gamma' \end{vmatrix}$ is nonzero. Since we are not interested in identifiability at a point, but rather in identifiability "almost everywhere" [Koopmans et al., 1950; Simon, 1953], we need not compute this determinant. We merely inspect the symbolic form of the determinant's rows to make sure that the equations are nonredundant; each imposes a new constraint on the unlabeled coefficients for at least one value of the labeled coefficients.

With a facility to detect redundancies, we can increase the power of our procedure by adding the following rule:

3b. If there are k nonredundant buckets that contain at most k unlabeled coefficients, label these coefficients, and continue.

Another way to increase the power of our procedure is to list not just identifiable effects, but also expressions involving correlations due to bi-directed arcs, in accordance with Wright's rules. Finally, one can endeavor to list effects of several variables jointly. A modified back-door criterion for evaluating joint effects has been reported by Pearl and Robins [1995]. However, such enrichments tend to make the procedure more complex and might compromise our main objective of providing investigators with a way of immediately recognizing the identified coefficients in a given model and immediately understanding those features in the model that influence the identifiability of the target quantity. We now address the problem of identification in nonparametric models, where the machinery of linear algebra can be of little help and where graph theoretical techniques have led to significant progress.

3.2 Identification in Nonparametric Models

Nonparametric models are structural equation models in which both the functional forms of the equations and the probability distributions of the disturbances remain unspecified. We consider nonparametric models for both practical and conceptual reasons. On the practical side, investigators often find it hard to defend the assumptions of linearity and normality, or other functional-distributional assumptions, especially when categorical variables are involved. Nonparametric results are valid for nonlinear functions and for any distribution of errors. Moreover, having such results allows us to gauge how sensitive standard techniques are to assumptions of linearity and normality. On the conceptual side, nonparametric models, which are stripped of algebraic connotations, illuminate the distinctions between structural and algebraic equations. The search for alternatives to path coefficients (which are nonexistent in nonparametric models) and to the standard definitions of direct and total causal effects (which are normally defined in terms of path coefficients) forces explication of what path coefficients really mean and of where their empirical content comes from.

3.2.1 Parametric vs. nonparametric models: An example

Consider the set of structural equations

$$\begin{aligned} x &= f_1(u, \epsilon_1) && (6) \\ z &= f_2(x, \epsilon_2) && (7) \\ y &= f_3(z, u, \epsilon_3) && (8) \end{aligned}$$

where X, Z, and Y are observed variables, f_1, f_2, and f_3 are unknown arbitrary functions, and $U, \epsilon_1, \epsilon_2$, and ϵ_3 are unobservables that we can regard either as latent variables or as disturbances. For the sake of this discussion, we will assume that $U, \epsilon_1, \epsilon_2$, and ϵ_3 are mutually independent and arbitrarily distributed. Graphically, these influences can be represented by the path diagram of Fig. 13.

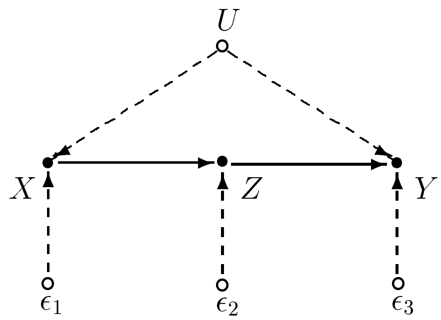

Figure 13: Path diagram corresponding to Eqs. (6)-(8), where $\{X, Z, Y\}$ are observed and $\{U, \epsilon_1, \epsilon_2, \epsilon_3\}$ are unobserved.

The problem is as follows: we have drawn a long stream of independent samples of the process defined by Eqs. (6)–(8) and have recorded the values of the observed variables X, Z, and Y, and we now wish to estimate the unspecified quantities of the model to the greatest extent possible.

To clarify the scope of the problem, let us consider its linear version, which is given by

$$\begin{aligned} x &= u + \epsilon_1 && (9) \\ z &= \alpha x + \epsilon_2 && (10) \\ y &= \beta z + \gamma u + \epsilon_3 && (11) \end{aligned}$$

where $U, \epsilon_1, \epsilon_2$, and ϵ_3 are uncorrelated, zero-mean disturbances.[15] It is not hard to show that parameters α, β, and γ can be determined uniquely from the correlations among the observed quantities X, Z, and Y. This identification was demonstrated already in the example of Fig. 8, where the back-door criterion yielded

$$\beta = r_{YZ \cdot X} \qquad \alpha = r_{ZX} \tag{12}$$

and hence

$$\gamma = r_{YX} - \alpha\beta \tag{13}$$

Thus, returning to the nonparametric version of the model, it is tempting to generalize that for the model to be identifiable, the functions $\{f_1, f_2, f_3\}$ must be determined uniquely from the data. However, the prospect of this happening is unlikely, because the mapping between functions and distributions is known to be many to one. In other words, given any nonparametric model M, if there exists one set of functions $\{f_1, f_2, f_3\}$ compatible with a given distribution $P(x, y, z)$, then there are infinitely many such functions. Thus, it seems that nothing useful can be inferred from loosely specified models such as the one given by Eqs. (6)–(8).

Identification is not an end in itself, however, even in linear models. Rather it serves to answer practical questions of prediction and control. At issue is not whether the data permit us to identify the form of the equations but rather whether the data permit us to provide unambiguous answers to questions of the kind traditionally answered by parametric models.

When the model given by Eqs. (6)–(8) is used strictly for prediction (i.e., to determine the probabilities of some variables given a set of observations on other variables), the structural content of the parameters becomes irrelevant; the predictions can be estimated directly from either the covariance matrices or the sample estimates of those covariances. If dimensionality reduction is needed (e.g., to improve estimation accuracy), the covariance matrix can be encoded in a variety of simultaneous equation models, all of the same dimensionality. For example, the correlations among X, Y, and Z in the linear model M of Eqs. (9)–(11) might well be represented by the model M' (Fig. 14):

$$\begin{aligned} x &= \epsilon_1 && (14)\\ z &= \alpha' x + \epsilon_2 && (15)\\ y &= \beta' z + \delta x + \epsilon_3 && (16) \end{aligned}$$

which is as compact as Eqs. (9)–(11) and is covariance equivalent to M with respect to the observed variables X, Y and Z. Upon setting $\alpha' = \alpha, \beta' = \beta$, and $\delta = \gamma$, model M' will yield the same probabilistic predictions as those of the model of Eqs. (9)–(11). Still, when viewed as data-generating mechanisms, the two models are not equivalent; each tells a different story about the processes generating X, Y, and Z, and, naturally, their predictions about the changes that would result from subjecting these processes to external interventions differ.

[15] An equivalent version of this model is obtained by eliminating U from the equations and allowing ϵ_1 and ϵ_3 to be correlated.

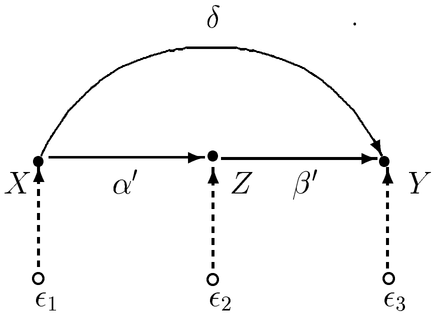

Figure 14: Diagram representing model M' of Eqs. (14)–(16).

3.3 Causal Effects: The Interventional Interpretation of Structural Equation Models

The differences between models M and M' illustrate precisely where the structural reading of simultaneous equation models comes into play. Model M', defined by Eqs. (14)–(16), regards X as a direct participant in the process that determines the value of Y, while model M, defined by Eqs. (9)–(11), views X as an indirect factor whose effect on Y is mediated by Z. This difference is not manifested in the data but in the way the data would change in response to outside interventions. For example, suppose we wish to predict the expectation of Y after we intervene and fix the value of X to some constant x, denoted $E(Y|do(X = x))$.[16] After $X = x$ is substituted into Eqs. (15) and (16), model M' yields

$$\begin{aligned} E[Y|do(X = x)] &= E[\beta'\alpha' x + \beta'\epsilon_2 + \delta x + \epsilon_3] && (17) \\ &= (\beta'\alpha' + \delta)x && (18) \end{aligned}$$

while model M yields

$$\begin{aligned} E[Y|do(X = x)] &= E[\beta\alpha x + \beta\epsilon_2 + \gamma u + \epsilon_3] && (19) \\ &= \beta\alpha x && (20) \end{aligned}$$

Upon setting $\alpha' = \alpha$, $\beta' = \beta$, and $\delta = \gamma$ (as required for covariance equivalence, see Eqs. (12) and (13)), we see clearly that the two models assign different magnitudes to the (total) causal effect of X on Y; model M predicts that a unit change in x will change $E(Y)$ by the amount $\beta\alpha$, while model M' puts this amount at $\beta\alpha + \gamma$.

At this point, it is tempting to ask whether we should substitute $x - \epsilon_1$ for u in Eq. (11) prior to taking expectations in Eq. (19). If we permit the substitution of Eq. (10) into Eq. (11), as we did in deriving Eq. (19), why not permit the substitution of Eq. (9) into Eq. (11) as well? After all, so the argument goes, there is no harm in upholding a mathematical equality, $u = x - \epsilon_1$, that the modeler deems valid. This argument is fallacious, however. Structural equations are not meant to be treated as immutable mathematical equalities. Rather, they are introduced into the model to describe a state of equilibrium, and they are *violated* when that equilibrium is perturbed by outside interventions. In fact, the power of

[16] Pearl [1993, 1995] used the notation $set(X = x)$. Currently, however, the $do(X = x)$ notation (taken from [Goldszmidt and Pearl, 1992] seems to be winning broader popular support.

structural equation models is that they not only encode the initial equilibrium state but also the information necessary for determining which equations must be violated to account for a new state of equilibrium. For example, if the intervention merely consists of holding X constant at x, then the equation $x = u + \epsilon_1$, which represents the pre-intervention process determining X, should be overruled and replaced with the equation $X = x$. The solution to the new set of equations then represents the new equilibrium. Thus, the essential characteristic of structural equations that sets them apart from ordinary mathematical equations is that they stand not for one but for many sets of equations, each corresponding to a subset of equations taken from the original model. Every such subset represents some hypothetical physical reality that would prevail under a given intervention.

If we take the stand that the value of structural equations lies not in summarizing distribution functions but in encoding causal information for predicting the effects of policies [Haavelmo, 1943; Marschak, 1950; Koopmans, 1953], it is natural to view such predictions as the proper generalization of structural coefficients. For example, the proper generalization of the coefficient β in the linear model M would be the answer to the control query, "What would be the change in the expected value of Y if we were to intervene and change the value of Z from z to $z+1$," which is different, of course, from the observational query, "What would be the difference in the expected value of Y if we were to *find* Z at level $z+1$ instead of level z." Observational queries can be answered directly from the joint distribution $P(x, y, z)$, while control queries require causal information as well. Structural equations encode this causal information in their syntax, by treating the variable on the left hand side of the equality sign as the effect and those on the right as causes. To distinguish between the two types of queries, we will use the symbol $do(\cdot)$ to indicate externally controlled quantities. For example, we write

$$E(Y|do(x)) \triangleq E[Y|do(X = x)] \tag{21}$$

for the controlled expectation and

$$E(Y|x) \triangleq E(Y|X = x) \tag{22}$$

for the standard conditional or observational expectation. That $E(Y|do(x)) \neq E(Y|x)$ can easily be seen in the model of Eqs. (9)–(11), where $E(Y|do(x)) = \alpha\beta x$ but $E(Y|x) = r_{YX}x = (\alpha\beta + \gamma)x$. Indeed, the passive observation $X = x$ should not violate any of the equations, and this is the justification for substituting both Eqs. (9) and (10) into Eq. (11) before taking the expectation.

In linear models, the answers to questions of direct control are encoded in the so-called path coefficients or structural coefficients, and these can be used to derive the total effect of any variable on another. For example, the value of $E(Y|do(x))$ in the model defined by Eqs. (9)–(11) is $\alpha\beta x$, namely, x times the product of the path coefficients along the path $X \longrightarrow Z \longrightarrow Y$. In the nonparametric case, computation of $E(Y|do(x))$ would be more complicated, even when we know the functions f_1, f_2, and f_3. Nevertheless, this computation is well defined and requires the solution (for the expectation of Y) of a modified set of equations in which f_1 is "wiped out" and X is replaced by the constant x:

$$z = f_2(x, \epsilon_2) \tag{23}$$

$$y = f_3(z, u, \epsilon_3) \tag{24}$$

Thus, computation of $E(Y|do(x))$ requires evaluation of

$$E(Y|do(x)) = E\{f_3[f_2(x, \epsilon_2), u, \epsilon_3]\}$$

where the expectation is taken over U, ϵ_2, and ϵ_3. Graphical methods for performing this computation are discussed in Section 3.4.

What, then, is an appropriate definition of identifiability for nonparametric models? One reasonable definition is that answers to interventional queries are *unique*. Accordingly, we call a model *identifiable* if there exists a consistent estimate for every query of the type "Find $P(r|do(s)) \triangleq P[R = r|do(S = s)]$," where R and S are subsets of observables and r and s are any realizations of these variables. The set of probabilities $P(r|do(s))$ is called the *causal effect* of S on R, as it describes how the distribution of R varies when S is changed by external control.[17] Naturally, we need to allow for instances in which some queries are identifiable while the system as a whole is not. Hence, we say that $P(r|do(s))$ is identifiable in model M if every choice of the model's parameters (i.e., functional forms and distributions) compatible with the observed distribution P yields the same value for $P(r|do(s))$.

Remarkably, many aspects of nonparametric identification, including tests for deciding whether a given interventional query is identifiable, as well as formulas for estimating such queries, can be determined graphically, almost by inspection, from the diagrams that accompany the equations.

3.4 Identification of Causal Effects

Definition 4 (causal effect) *Given a causal model M (as in Eq. (1)) and two disjoint sets of variables, X and Y, the* causal effect *of the set X on the set Y, denoted $P_M(y|do(x))$, is the probability of $Y = y$ induced by deleting from the model all equations corresponding to variables in X and substituting $X = x$ in the remaining equations.*[18]

Clearly, the graph corresponding to the reduced set of equations is an edge subgraph of G from which all arrows to X have been pruned [Spirtes et al., 1993].

Readers accustomed to thinking of causal effects in terms of randomized experiments may interpret $P(y|do(x))$ as the conditional probability $P_{exp}(Y = y|X = x)$ corresponding to a controlled experiment in which X is randomized. An equivalent interpretation can be formulated using the potential response notation [Rubin, 1974] to read

$$P(y|do(x)) = P(Y_x = y)$$

where Y_x is the value that Y would obtain under the hypothetical control (or treatment) $do(X = x)$. Rubin's definition of causal effect, $E(Y_{x'}) - E(Y_{x''})$, where x' and x'' are two levels of a treatment variable X, corresponds to the difference $E(y|do(x')) - E(Y|do(x''))$,

[17] Technically, the adjective "causal" is redundant. It serves to emphasize, however, that the changes in S are enforced by external control and do not represent stochastic variations in the observed value of S.

[18] Explicit translation of interventions to "wiping out" equations from the model was first proposed by Strotz and Wold [1960] and has since been used by Fisher [1970] and Sobel [1990]. Graphical ramifications of this translation were explicated by Spirtes et al. [1993] and Pearl [1993]. A related mathematical model using event trees has been introduced by Robins [1986, pp. 1422–1425].

and can always be obtained from the generic distribution $P(y|do(x))$. Definition 4 forms the bridge between structural equation models and the potential response framework. It provides a precise model-theoretic definition for the counterfactual variable Y_x, which in the potential response framework is taken as a hypothetical mental construct. The SEM equivalent of Y_x is: *the solution for Y, after deleting from the model all equations corresponding to variables in the set X, and substituting $X = x$ in the remaining equations.*

Definition 5 (causal effect identifiability) *The causal effect of X on Y is said to be identifiable in a class C of models if the quantity $P(y|do(x))$ can be computed uniquely from the probabilities of the observed variables V, that is, if for every pair of models M_1 and M_2 in C for which $P_{M_1}(v) = P_{M_2}(v)$, we have $P_{M_1}(y|do(x)) = P_{M_2}(y|do(x))$.*

Our analysis of identifiability will focus on a class C of models that have the following characteristics in common:

1. they share the same causal graph G, and
2. they induce positive distributions on the observed variables, that is, $P_M(v) > 0$.

Analysis of causal effects becomes particularly simple when dealing with Markovian models. In such models, all causal effect queries are identifiable, that is, they can be computed directly from the conditional probabilities $P(x_i|pa_i)$, even when the functional forms of the functions f_i and the distributions of the disturbances are not specified [Pearl, 1993c; Spirtes et al., 1993]. This is seen immediately from the following observations. On the one hand, the distribution induced by any Markovian model M is given by the product in Eq. (3),

$$P_M(x_1, \ldots, x_n) = \prod_i P(x_i|pa_i) \tag{25}$$

where pa_i are (values of) the parents of X_i in the diagram representing M. On the other hand, the submodel $M_{x'_j}$, representing the action $do(X_j = x'_j)$, is also Markovian; hence, it also induces a product-like distribution

$$P_{M_{x'_j}}(x_1, \ldots, x_n) \; = \begin{cases} \prod_{i \neq j} P(x_i|pa_i) = \frac{P(x_1,\ldots,x_n)}{P(x_j|pa_j)} & \text{if } \; x_j = x'_j \\ 0 & \text{if } \; x_j \neq x'_j \end{cases} \tag{26}$$

where the partial product reflects the removal of the equation $x_j = f_j(pa_j, \epsilon_j)$ from the model. Thus, we see that both the pre-action and the post-action distributions depend only on observed conditional probabilities, but they are independent of the particular functional forms of $\{f_i\}$ and of the error distributions that generated those probabilities.

It is possible to show that certain, although not all, causal effects are identifiable in semi-Markovian nonparametric models [Pearl, 1995]. An important result in this direction has been the extension of the back-door criterion (Theorem 7) to nonparametric models:

Theorem 8 (nonparametric back-door criterion) *For any disjoint sets of variables X and Y in a causal diagram G, if the two conditions of Theorem 7 are satisfied, then the causal effect of X on Y is identified and is given by*

$$P(y|do(x)) = \sum_z P(y|x, z)P(z) \tag{27}$$

This theorem provides a formal definition for the concepts of *exogeneity* and *confounding* in econometrics [Engle et al., 1983] and epidemiology [Greenland et al., 1998], respectively. A variable X is said to be exogenous (unconfounded) relative to Y if $P(y|do(x)) = P(y|x)$, that is, if the conditions of the back-door criterion hold when Z is the empty set. Alternatively, X is said to be conditionally exogenous (unconfounded) relative to Y, given measurements on set Z, if Eq. (27) holds, that is, if the conditions of the back-door criterion hold for Z. Section 3.1 proposes an explanation why the definitions of these two basic concepts have encountered difficulties in econometrics and epidemiology (see also [Pearl, 1998]).

Pearl [1995] introduces a symbolic calculus for the $do(\cdot)$ operator, which facilitates the identification of additional causal effects in nonparametric models. Using this calculus, Galles and Pearl [1995] have devised graphical criterion for identifying causal effects in any semi-Markovian model. Finally, if the objective of a study is to evaluate the direct, rather than the total, causal effect of X on Y, as was the case with the Berkeley graduate admissions study (see footnote 5), then some other graphical criteria that determine identifiability are available [Pearl and Robins, 1995]. Applications to policy analysis and to the management of noncopliance are presented in [Balke and Pearl, 1995, 1997].

In light of these results, the reader might want to know whether the model defined by Eqs. (6)–(8) is identifiable. The answer is yes; this model permits the identification of all interventional queries. For example, from inspection of the graph in Fig. 13, we can conclude immediately that

1. $P(x|do(y), do(z)) = P(x)$,
   consistent with the intuition that consequences can have no effect on their causes;

2. $P(z|do(x)) = P(z|x)$,
   because ϵ_2 is independent of X, and hence Z is not confounded by X
   (alternatively, and hence all back-door paths between Z and X are blocked);

3. $P(y|do(z)) = \sum_x P(y|z,x)P(x)$,
   because the back-door criterion qualifies X as an appropriate covariate for adjustment; and

4. $P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x',z)P(x')$,
   which results from chaining $P(z|do(x))$ with $P(y|do(z))$, as is shown formally in [Pearl, 1995].

4 SOME CONCEPTUAL UNDERPINNINGS

4.1 What Do Structural Parameters Really Mean?

Every student of SEM has stumbled on the following paradox at some point in his or her career. If we interpret the coefficient β in the equation

$$y = \beta x + \epsilon$$

as the change in $E(Y)$ per unit change of X, then after rewriting the equation as

$$x = (y - \epsilon)/\beta$$

we ought to interpret $1/\beta$ as the change in $E(X)$ per unit change of Y. But this conflicts both with intuition and with the prediction of the model: the change in $E(X)$ per unit change of Y ought to be zero if Y does not appear as an independent variable in the equation for X.

Teachers of SEM generally evade this dilemma via one of two escape routes. One route involves denying that β has any causal reading and settling for a purely statistical interpretation in which β measures the reduction in the variance of Y explained by X (e.g., [Muthen, 1987]). The other route permits causal reading of only those coefficients that meet the so-called isolation restriction [Bollen, 1989; James et al., 1982], namely, the explanatory variable must be uncorrelated with the error in the equation. Since ϵ cannot be uncorrelated with both X and Y, so the argument goes, β and $1/\beta$ cannot both have causal meaning, and the paradox dissolves.

The first route is self-consistent, but it compromises the founders' intent that SEM function as an aid to policy making and clashes with the intuition of most SEM users. The second is vulnerable to attack logically. It is well known that every pair of bi-variate normal variables, X and Y, can be expressed in two equivalent ways:

$$y = \beta x + \epsilon_1$$

and

$$x = \alpha y + \epsilon_2$$

where $cov(X, \epsilon_1) = cov(Y, \epsilon_2) = 0$, and $\alpha = r_{XY} = \beta\sigma_X^2/\sigma_Y^2$. Thus, if the condition $cov(X, \epsilon_1) = 0$ endows β with causal meaning, then $cov(Y, \epsilon_2) = 0$ ought to endow α with causal meaning as well. But this, too, conflicts with both intuition and the intentions behind SEM; the change in $E(X)$ per unit change of Y ought to be zero, not r_{XY}, if there is no arrow from Y to X.

What then *is* the meaning of a structural coefficient? Or a structural equation? Or an error term? The interventional interpretation of causal effects, when coupled with the $do(x)$ notation introduced in Section 3.3, provides simple answers to these questions. The answers explicate the operational meaning of structural equations and thus should, I hope, end an era of controversy and confusion regarding these entities.

4.1.1 Structural equations: Operational definition

Definition 6 (structural equations) *An equation $y = \beta x + \epsilon$ is said to be* structural *if it is to be interpreted as follows: In an ideal experiment where we control X to x and any other set Z of variables (not containing X or Y) to z, the value y of Y would be independent of z and is given by $\beta x + \epsilon$.*

This definition is operational because all quantities are observable, albeit under conditions of controlled manipulation. That manipulations cannot be performed in most observational studies does not negate the operationality of the definition, in much the same way that our inability to observe bacteria with the naked eye does not negate their observability under a microscope. The challenge of SEM is to extract the maximum information on what we wish to observe, from the little we can observe.

Note that the operational reading given above makes no claim about how X (or any other variable) will behave when we control Y. This asymmetry makes the equality signs in structural equations different from algebraic equality signs; the former act symmetrically in relating observations on X and Y (e.g., observing $Y = 0$ implies $\beta x = -\epsilon$), but they act asymmetrically when it comes to interventions (e.g., setting Y to zero tells us nothing about the relation between x and ϵ). The arrows in path diagrams make this dual role explicit, and this may account for the insight and inferential power gained through the use of diagrams.

The strongest empirical claim of the equation $y = \beta x + \epsilon$ is made by excluding other variables from the r.h.s. of the equation, thus proclaiming X the *only* immediate cause of Y. This translates into a testable claim of *invariance*: the statistics of Y under condition $do(x)$ should remain invariant to the manipulation of any other variable in the model [Pearl, 1996; Galles and Pearl 1998].[19] This can be written symbolically as

$$P(y|do(x), do(z)) = P(y|do(x)) \tag{28}$$

for all Z disjoint of $\{X \cup Y\}$.

Note that this invariance holds relative to manipulations, not observations, of Z. The statistics of Y under condition $do(x)$, given the measurement $Z = z$, written $P(y|do(x), z)$, would certainly depend on z if the measurement was taken on a consequence (i.e., descendant) of Y. Note also that the ordinary conditional probability $P(y|x)$ does not enjoy such a strong property of invariance, since $P(y|x)$ is generally sensitive to manipulations of variables other than X in the model, unless X and ϵ are independent. Eq. (28), in contrast, remains valid regardless of the statistical relationship between ϵ and X.

Generalized to a set of several structural equations, Eq. (28) explicates the assumptions underlying a given causal diagram. If G is the graph associated with a set of structural equations, then the assumptions are embodied in G as follows: (1) every missing arrow, say between X and Y, represents the assumption that X has no causal effect on Y once we intervene and hold the parents of Y fixed; and (2) every missing bi-directed link between X and Y represents the assumption that the omitted factors that influence Y are uncorrelated with those that influence X. Sections 4.1.3 and 4.1.4 discuss the operational meaning of the latter assumption.

4.1.2 The structural parameters: Operational definition

The interpretation of a structural equation as a statement about the behavior of Y under a hypothetical intervention yields a simple definition for the structural parameters. The meaning of β in the equation $y = \beta x + \epsilon$ is simply

$$\beta = \frac{\partial}{\partial x} E[Y|do(x)] \tag{29}$$

namely, the rate of change (relative to x) of the expectation of Y in an experiment where X is held at x by external control. This interpretation holds regardless of whether ϵ and X are correlated in nonexperimental studies (e.g., via another equation $x = \alpha y + \delta$).

[19]This statistical claim is in fact only part of the message conveyed by the equation; the other part consists of a counterfactual claim [Lewis, 1973; Rubin, 1990; Galles and Pearl, 1998]: if we were to control X to x', instead of x, then Y would attain the value $\beta x' + \epsilon$. In other words, plotting the value of Y under various hypothetical controls of X, and under the same external conditions (ϵ), should result in a straight line with slope β. Such claims can only be tested under the assumption that ϵ, representing external conditions or properties of experimental units remains unaltered as we switch from x to x' [Pearl, 1996, p. 404].

4.1.3 The mystical error term: Operational definition

The interpretations given in Sections 4.1.1 and 4.1.2 provide an operational definition for that mystical error term

$$\epsilon = y - E[Y|do(x)] \tag{30}$$

which, despite being unobserved in nonmanipulative studies is far from being metaphysical or definitional as suggested by some researchers (e.g., [Richard, 1980; Holland, 1988, p. 460; Hendry, 1995, p. 62]). Unlike errors in regression equations, ϵ measures the deviation of Y from its controlled expectation $E[Y|do(x)]$ and not from its conditional expectation $E[Y|x]$. The statistics of ϵ can therefore be measured from observations on Y once X is controlled. Alternatively, since β remains the same, regardless of whether X is manipulated or observed, the statistics of $\epsilon = y - \beta x$ can be measured in observational studies, if we know β.

Likewise, correlations among errors can be estimated empirically. For any two nonadjacent variables, X and Y, Eq. (30) yields:

$$E[\epsilon_Y \epsilon_X] = E[YX|do(pa_Y, pa_X)] - E[Y|do(pa_Y)]E[X|do(pa_X)]\} \tag{31}$$

Once we determine the structural coefficients, the controlled expectations $E[Y|do(pa_Y)]$, $E[X|do(pa_X)]$, and $E[YX|do(pa_Y, pa_X)]$ become known linear functions of the observed variables pa_Y and pa_X, hence the expectations on the r.h.s. of Eq. (31) can be estimated in observational studies. Alternatively, if the coefficients are not determined, the expression can be assessed directly in interventional studies, by holding pa_X and pa_Y fixed (assuming X and Y are not in parent-child relationship) and estimating the covariance of X and Y from data obtained under such conditions.

Finally, we are often interested not in assessing the numerical value of $E[\epsilon_Y \epsilon_X]$ but rather in determining whether ϵ_Y and ϵ_X can be assumed uncorrelated. For this determination, it suffices to test whether the equality

$$E[Y|x, do(s_{XY})] = E[Y|do(x), do(s_{XY})] \tag{32}$$

holds true, where s_{XY} stands for (any setting of) all variables in the model excluding X and Y. This test can be applied to any two variables in the model, except when Y is a parent of X, in which case the symmetrical equation is applicable, with X and Y interchanged.

4.1.4 The mystical error term: Conceptual interpretation

The authors of SEM textbooks usually interpret error terms as representing the influence of omitted factors. Many SEM researchers are reluctant to accept this interpretation, however, partly because unspecified omitted factors open the door to metaphysical speculations, and partly because arguments based on such factors were improperly used as a generic, substance-free license to omit bi-directed arcs from path diagrams [McDonald, 1997]. Such concerns are answered by the operational interpretation of error terms, Eq. (30), since it prescribes how errors are measured, not how they originate.

It is important to note, though, that the operational definition is no substitute for the omitted-factors conception when it comes to deciding whether pairs of error terms can be assumed to be uncorrelated. Since such decisions are needed at a stage when the model's parameters are still "free," they cannot be made on the basis of numerical assessments of correlations but must rest instead on qualitative structural knowledge about how mechanisms are tied together and how variables affect each other. Such judgmental decisions are hardly aided by the operational criterion of Eq. (31), because that criterion instructs the investigator to assess whether two deviations, taken on two different variables under complex experimental conditions, would be correlated or uncorrelated. Such assessments are cognitively unfeasible.

In contrast, the omitted-factors conception instructs the investigator to judge whether there could be factors that simultaneously influence several observed variables. Such judgments are cognitively manageable, because they are qualitative and rest on purely structural knowledge – the only knowledge available during this phase of modeling.

Another source of error correlation is *selection bias*. If two uncorrelated unobserved factors have a common effect that is omitted from the analysis but influences the selection of samples for the study, then the corresponding error terms will be correlated in the sampled population, that is, the expectation in Eq. (31) will not vanish when taken over the sampled population (see discussion of Berkson's paradox following Definition 1).

We should emphasize, however, that the arcs *missing* from the diagram, not those *in* the diagram, demand the most attention and careful substantive justification. Adding an extra bi-directed arc can only compromise the identifiability of parameters, but deleting an existing bi-directed arc may produce erroneous conclusions together with a false sense of model testability. Thus, bi-directed arcs should be assumed to exist, by default, between any two nodes in the diagram. They should be deleted only by well-motivated justifications, such as the unlikely existence of a common cause for the two variables and the unlikely existence of selection bias. Although we can never be cognizant of all the factors that may affect our variables, substantive knowledge sometimes permits us to state that if a common factor exists, its influence is not likely to be significant.

Thus, as often happens in the sciences, the way we measure physical entities does not offer the best way of thinking about them. The omitted-factor conception of errors, because it rests on structural knowledge, is a more useful guide than the operational definition when building, evaluating, and thinking about causal models.

4.2 Interventional Interpretation of Effect Decomposition

In this section, we show that the interventional interpretation of structural equations provides simple, unambiguous definitions of effect decomposition for both parametric and nonparametric models.

We start with the general notion of a causal effect $P(y|do(x))$ from Definition 4, which applies to arbitrary sets of variables, X and Y. This interpretation of causal effect can be specialized to define total and direct causal effects, as follows.

Definition 7 (total effect) *The total effect of X on Y is given by $P(y|do(x))$, namely, the distribution of Y while X is held constant at x and all other variables are permitted to run their natural course.*

Definition 8 (direct effect) *The direct effect of X on Y is given by $P(y|do(x),\ do(s_{XY}))$ where S_{XY} is the set of all observed variables except X and Y in the system.*

In linear analysis, Definition 7 and 8 yield, after differentiation w.r.t. x, the path coefficients that are normally associated with direct and indirect effects, yet they differ from conventional definitions in several important aspects. First, direct effects are defined in terms of hypothetical experiments in which intermediate variables are held constant by *physical intervention*, not by statistical adjustment (which is often disguised under the misleading phrase "control for"). Fig. 11 depicts a simple example where adjustmenting for the intermediate variables (Z and W) would not give the correct value (= 0) for the direct effect of X on Y, whereas $\frac{\partial}{\partial_x} E(Y|do(x, z, w))$ does yield the correct value: $\frac{\partial}{\partial_x}(\beta w + \gamma z) = 0$. Second, there is no need to limit control to only intermediate variables; *all* variables in the system may he held constant (excluding X and Y). Thus, the measurement of direct effects is ascribed to an ideal laboratory; the scientist controls for all possible conditions S_{XY}, and measurements may commence without knowing the structure of the diagram. Finally, our definitions differ from convention by interpreting total and direct effects independently of each other, as outcomes of two different experiments. Textbook definitions (e.g., [Bollen, 1989, p. 376]) usually equate the total effect with a power series of path coefficient matrices. This algebraic definition coincides with the operational definition (Definition 7) in recursive (semi-Markovian) systems, but yields erroneous expressions in models with feedback. For instance, given the pair of equations $\{y = \beta x + \epsilon,\ x = \alpha y + \delta\}$, the total effect of X on Y is simply β, not $\beta(1 - \alpha\beta)^{-1}$ as stated in [Bollen, 1989, p. 379]. The latter has no operational significance worthy of the phrase "effect of X."[20]

Note that when the structure of the causal diagram is known, there is no need to actually hold *all* other variables constant; holding constant the direct parents of Y (excluding X) would have the same effect on Y as holding all variables except X and Y constant. Thus, we obtain the following equivalent definition of a direct effect.

Corollary 2 *The direct effect of X on Y is given by $P(y|do(x),\ do(pa_{Y\backslash X}))$, where $pa_{Y\backslash X}$ stands for any realization of the variables appearing in the equation for Y, excluding X.*

Readers versed in linear analysis may find it a bit strange that the direct effect of X on Y involves variables other than X and Y. However, given that we are dealing with nonlinear interactions, the effect of X on Y should indeed depend on the levels at which we hold the other parents of Y. If we wish to average over these values and take the expectation of Y, we obtain the expression

$$\Delta_{x,x'}(Y) = \sum_{pa_{Y\backslash X}} [E(Y|do(x'), do(pa_{Y\backslash X})) - E(Y|do(x), do(pa_{Y\backslash X}))] P(pa_{Y\backslash X}|do(x))$$

where $\Delta_{x,x'}(Y)$ stands for the average change in $E(Y)$ induced by changing X from x to x' while keeping the other parents of Y constant at whatever value they obtain under $do(x)$. This expression explicates what we actually wish to measure in race or sex discrimination cases, where we are instructed to assess the effect of one factor (X) while keeping "all other factors constant". If X is an exogenous variable, as "gender" is in Berkeley's sex discrimination case (see footnote 5), then $do(x)$ can be replaced with x. However, we are

[20] This error was noted by Sobel (1990) but, perhaps because constancy of path coefficients was presented as new and extraneous assumptions, Sobel's correction has not brought about a shift in practice or philosophy.

not at liberty to replace $do(pa_{Y\setminus X})$ with $pa_{Y\setminus X}$ unless we can safely assume that the factors represented by $pa_{Y\setminus X}$ (e.g., student's qualifications and the choice of department) are not confounded with Y (as defined in Theorem 8). In general, we see that the average direct effect $\Delta_{x,x'}(Y)$ is identifiable whenever both $E[Y|do(x), do(pa_{Y\setminus X})]$ and $P(pa_{Y\setminus X}|do(x))$ are identifiable. Note that if X does not appear in the equation for Y, then $P(y|do(x), do(pa_{Y\setminus X}))$ defines a constant distribution on Y, independent of x, which matches our understanding of "having no direct effect." Note as well that, in linear models, $\Delta_{x,x'}(Y)$ reduces to $x' - x$ times the path coefficient between X and Y.

In standard linear analysis, an indirect effect may be defined as the difference between the total and the direct effects [Bollen, 1989]. In nonlinear analysis, differences lose their significance, and one must isolate the contribution of mediating paths in some other way. Expressions of the form $P(y|do(x), do(z))$ cannot be used to isolate such contributions, however, because there is no physical means of selectively disabling a direct causal link from X to Y by holding some variables constant. This suggests that the notion of indirect effect has no intrinsic operational meaning apart from providing a comparison between the direct and the total effects. In other words, a policymaker who asks for that part of the total effect transmitted by a particular intermediate variable or by a group Z of such variables is really asking for a comparison of the effects of two policies, one where Z is held constant, the other where it is not. The expressions corresponding to these policies are $P(y|do(x),\ do(z))$ and $P(y|do(x))$, and this pair of distributions should be taken as the most general representation of indirect effects. Similar conclusions have been expressed by Robins [1986] and Robins and Greenland [1992].

5 CONCLUSION

Today the enterprise known as structural equation modeling (SEM) is increasingly under fire. The founding fathers have retired, their teachings are forgotten, and practitioners, teachers, and researchers currently find the methodology they inherited difficult to either defend or supplant. Modern SEM textbooks are preoccupied with parameter estimation and rarely explicate the role that those parameters play in causal explanations or in policy analysis; examples dealing with the effects of interventions are conspicuously absent, for instance. Research in SEM now focuses almost exclusively on model fitting, while foundational issues pertaining to the meaning of SEM's models are subjects of confusion and controversy.

The contemporary crisis in SEM originates, I am thoroughly convinced, in the lack of a mathematical language that can handle the causal information embedded in structural equations. Recently, graphical models have provided such a language. They have thus helped us answer many of the unsettled foundational questions that drive the current crisis, including:

1. Under what conditions can we give causal interpretation to structural coefficients?
2. What are the causal assumptions underlying a given structural equation model?
3. What are the statistical implications of any given structural equation model?
4. What is the operational meaning of a given structural coefficient?

5. What are the policy-making claims of any given structural equation model?

6. When is an equation not structural?

In this paper, I have summarized the conceptual developments that now resolve such foundational questions. In addition, by way of illustrating the soundness of the proposed approach, I have presented several tools to be used in answering questions of practical importance. Questions of this type include:

1. When are two structural equation models observationally indistinguishable?

2. When do regression coefficients represent path coefficients?

3. When would the addition of a regressor introduce bias?

4. How can we tell, prior to taking any data, which path coefficients can be identified?

5. When can we dispose of the linearity-normality assumption and still extract causal information from the data?

I am hopeful that researchers will recognize the benefits of these concepts and tools and use them to revitalize causal analysis in the social and behavioral sciences.

Acknowledgments

This paper owes its inspiration to the generations of statisticians who have asked, with humor and disbelief, how SEM's methodology could make sense to any rational being, and to the social scientists who, perhaps unwittingly, have saved the SEM tradition from drowning in statistical interpretations. I am grateful to two (anonymous) referees for their exceptionally detailed comments on the original version of this paper and for correcting several oversights and omissions. The comments of Herman Ader, Jacques Hagenaars, and Rod McDonald have helped me gain a greater understanding of SEM practice and vocabulary. Jin Tiang was instrumental in revising Sections 2.3.1 and 3.1. This investigation has also benefited from discussions with John Aldrich, Peter Bentler, Nancy Cartwright, David Chickering, David Freedman, Arthur Goldberger, Sander Greenland, Scott Hershberger, Jim Heckman, Paul Holland, Guido Imbens, Jan Koster, Ed Leamer, Christopher Meek, Thomas Richardson, Jamie Robins, William Shipley, Peter Spirtes, and Rolf Steyer. This research was partially supported by grants from AFOSR, NSF, Northrop, and Rockwell.

References

[Andersson et al., 1998] S.A. Andersson, D. Madigan, M.D. Perlman, and T.S. Richardson. Graphical markov models in multivariate analysis. In Subir Ghosh, editor, *Multivariate analysis, Design of Experiments, and Survey Sampling*, page ??? Marcel Dekker, Inc., ?, 1998.

[Bagozzi and Burnkrant, 1979] R.P. Bagozzi and R.E. Burnkrant. Attitude organization and the attitude-behavior relationship. *Journal of Personality and Social Psychology*, 37:913–929, 1979.

[Balke and Pearl, 1995] A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. Morgan Kaufmann, San Francisco, 1995.

[Balke and Pearl, 1997] A. Balke and J. Pearl. Nonparametric bounds on causal effects from partial compliance data. *Journal of the American Statistical Association*, 92(439):1–6, September 1997.

[Berkson, 1946] J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2:47–53, 1946.

[Bickel et al., 1975] P.J. Bickel, E.A. Hammel, and J.W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–404, 1975.

[Blalock, Jr., 1962] H.M. Blalock, Jr. Four-variable causal models and partial correlations. *American Journal of Sociology*, 68:182–194, 1962.

[Bollen, 1989] K.A. Bollen. *Structural Equations with Latent Variables.* John Wiley, New York, 1989.

[Bowden and Turkington, 1984] R.J. Bowden and D.A. Turkington. *Instrumental Variables.* Cambridge University Press, Cambridge, England, 1984.

[Breckler, 1990] S.J. Breckler. Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin*, 107:260–273, 1990.

[Cartwright, 1995] N. Cartwright. Probabilities and experiments. *Journal of Econometrics*, 67:47–59, 1995.

[Chickering, 1995] D.M. Chickering. A transformational characterization of Bayesian network structures. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 87–98. Morgan Kaufmann, San Francisco, 1995.

[Chou and Bentler, 1995] C.P. Chou and P. Bentler. Estimations and tests in structural equation modeling. In R.H. Hoyle, editor, *Structural Equation Modeling*, pages 37–55. Sage, Thousand Oaks, CA, 1995.

[Cox and Wermuth, 1996] D.R. Cox and N. Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation.* Chapman and Hall, London, 1996.

[Dawid, 1979] A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series A*, 41:1–31, 1979.

[Duncan, 1975] O.D. Duncan. *Introduction to Structural Equation Models.* Academic Press, New York, 1975.

[Engle et al., 1983] R.F. Engle, D.F. Hendry, and J.F. Richard. Exogeneity. *Econometrica*, 51:277–304, 1983.

[Epstein, 1987] R.J. Epstein. *A History of Econometrics*. Elsevier Science, New York, 1987.

[Fisher, 1970] F.M. Fisher. A correspondence principle for simultaneous equations models. *Econometrica*, 38:73–92, 1970.

[Freedman, 1987] D. Freedman. As others see us: A case study in path analysis (with discussion). *Journal of Educational Statistics*, 12:101–223, 1987.

[Galles and Pearl, 1995] D. Galles and J. Pearl. Testing identifiability of causal effects. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 185–195. Morgan Kaufmann, San Francisco, 1995.

[Galles and Pearl, 1997] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97:9–43, 1997.

[Galles and Pearl, 1998] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. Technical Report R-250, Department of Computer Science, University of California, Los Angeles, 1998. To appear in *Foundations of Science*, 1998.

[Geiger et al., 1990] D. Geiger, T.S. Verma, and J. Pearl. Identifying independence in Bayesian networks. In *Networks*, volume 20, pages 507–534. John Wiley, Sussex, England, 1990.

[Goldberger, 1972] A.S. Goldberger. Structural equation models in the social sciences. *Econometrica*, 40:979–1001, 1972.

[Goldberger, 1992] A.S. Goldberger. Models of substance; comment on N. Wermuth, 'on block-recursive linear regression equations. *Brazilian Journal of Probability and Statistics*, 6:1–56, 1992.

[Goldszmidt and Pearl, 1992] M. Goldszmidt and J. Pearl. Stratified rankings for causal modeling and reasoning about actions. In *Proceedings of the Fourth International Workshop on Nonmonotonic Reasoning*, pages 99–110, Vermont, 1992.

[Grayson, 1987] D.A. Grayson. Confounding confounding. *American Journal of Epidemiology*, 126:546–553, 1987.

[Greenland and Robins, 1986] S. Greenland and J. Robins. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15:413–419, 1986.

[Greenland et al., 1998] S. Greenland, J. Pearl, and J.M Robins. Causal diagrams for epidemiologic research. Technical Report R-251, University of California, Los Angeles, January 1998. To appear in *Epidemiology*.

[Haavelmo, 1943] T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12, 1943.

[Hendry, 1995] David F. Hendry. *Dynamic Econometrics.* Oxford University Press, New York, 1995.

[Holland and Rubin, 1983] P.W. Holland and D.B. Rubin. On Lord's paradox. In H. Wainer and S. Messick, editors, *Principals of Modern Psychological Measurement*, pages 3–25. Lawrence Earlbaum, Hillsdale, NJ, 1983.

[Holland, 1986] P.W. Holland. Statistics and causal inference. *JASA*, 81:945–960, 1986.

[Holland, 1988] P.W. Holland. Causal inference, path analysis, and recursive structural equations models. In C. Clogg, editor, *Sociological Methodology*, pages 449–484. American Sociological Association, Washington, DC, 1988.

[Holland, 1995] P.W. Holland. Some reflections on Freedman's critiques. *Foundations of Science*, 1:50–57, 1995.

[James et al., 1982] L.R. James, S.A. Mulaik, and J.M. Brett. *Causal Analysis: Assumptions, Models, and Data.* Sage, Beverly Hills, 1982.

[Kim and Pearl, 1983] J.H. Kim and J. Pearl. A computational model for combined causal and diagnostic reasoning in inference systems. In A. Bundy, editor, *IJCAI-83: Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pages 190–193. William Kaufmann, Inc., Los Altos, CA, 1983.

[Koopmans et al., 1950] T.C. Koopmans, H. Rubin, and R.B. Leipnik. Measuring the equation systems of dynamic economics. In T.C. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, pages 53–237. John Wiley, New York, 1950.

[Koopmans, 1953] T.C. Koopmans. Identification problems in econometric model construction. In W.C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 27–48. Wiley, New York, 1953.

[Koster, 1998] J.T.A. Koster. On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. Technical Report EUR/FSW/97.03.02, Department of Sociology, Erasmus University Rotterdam, 1998.

[Lauritzen, 1996] S.L. Lauritzen. *Graphical Models.* Clarendon Press, Oxford, 1996.

[Leamer, 1985] E.E. Leamer. Vector autoregressions for causal inference? *Carnegie-Rochester Conference Series on Public Policy*, 22:255–304, 1985.

[Lee and Hershberger, 1990] S. Lee and S.A Hershberger. A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research*, 25:313–334, 1990.

[LeRoy, 1996] S.F. LeRoy. Causal orderings. In K.D. Hoover, editor, *Macroeconometrics: Developments, Tensions, Prospects*, pages 211–228. Kluwer Academic, Boston, 1996.

[Lewis, 1973] D. Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973. Reprinted with postscript in D. Lewis, *Philosophical Papers*, vol. 2. Oxford University Press, New York, 1986.

[Lucas Jr., 1976] R.E. Lucas Jr. Econometric policy evaluation: a critique. In Karl Brunner and Allan H. Meltzer (Eds.), editors, *The Phillips Curve and Labor Markets*, volume CRCS, Vol. 1, pages 19–46. North-Holland, Amsterdam, 1976.

[MacCallum et al., 1993] R.C. MacCallum, D.T. Wegener, B.N. Uchino, and L.R. Fabrigar. The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, 114:185–199, 1993.

[Marschak, 1950] J. Marschak. Statistical inference in economics. In T. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, pages 1–50. Wiley, New York, 1950.

[McDonald, 1997] R.P. McDonald. Haldane's lungs: A case study in path analysis. *Multivariate Behavioral Research*, 32:1–38, 1997.

[Meek, 1995] C. Meek. Causal inference and causal explanation with background knowledge. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 403–410. Morgan Kaufmann, San Francisco, 1995.

[Muthen, 1987] B. Muthen. Response to Freedman's critique of path analysis: Improve credibility by better methodological training. *Journal of Educational Statistics*, 12:178–184, 1987.

[Neyman,] J. Neyman. Sur les applications de la thar des probabilities aux experiences Agaricales: Essay de principle. English translation of excerpts by D. Dabrowska and T. Speed, *Statistical Science*, 5(4): 465–480, 1990.

[Pearl and Meshkat, 1998] J. Pearl and P. Meshkat. On testing regression models with fewer regressors. Technical Report R-257, Department of Computer Science, University of California, Los Angeles, 1998.

[Pearl and Robins, 1995] J. Pearl and J.M. Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 444–453. Morgan Kaufmann, San Francisco, 1995.

[Pearl and Verma, 1987] J. Pearl and T. Verma. The logic of representing dependencies by directed acyclic graphs. In *Proc., 6th National Conference on AI (AAAI-87)*, pages 374–379, Seattle, WA, July 1987.

[Pearl and Verma, 1991] J. Pearl and T. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452. Morgan Kaufmann, San Mateo, CA, 1991.

[Pearl, 1993a] J. Pearl. Belief networks revisited. *Artificial Intelligence*, 59:49–56, 1993.

[Pearl, 1993b] J. Pearl. Comment: Graphical models, causality and intervention. *Statistical Science*, 8:266–269, 1993.

[Pearl, 1993c] J. Pearl. From Bayesian networks to causal networks. In *Proceedings of the Adaptive Computing and Information Processing Seminar*, pages 25–27, Brunel Conference Centre, London, 1993. See also *Statistical Science*, 8: 266–269, 1993.

[Pearl, 1995] J. Pearl. Causal diagrams for experimental research. *Biometrika*, 82:669–710, 1995.

[Pearl, 1996] J. Pearl. Structural and probabilistic causality. In D.R. Shanks, K.J. Holyoak, and D.L. Medin, editors, *The Psychology of Learning and Motivation*, pages 393–435. Academic Press, San Diego, CA, 1996.

[Pearl, 1998] J. Pearl. Why there is no statistical test for confounding, why many think there is, and why they are almost right. Technical Report R-256, Department of Computer Science, University of California, Los Angeles, 1998.

[Pratt and Schlaifer, 1988] J.W. Pratt and R. Schlaifer. On the interpretation and observation of laws. *Journal of Econometrics*, 39:23–52, 1988.

[Richard, 1980] J.F. Richard. Models with several regimes and changes in exogeneity. *Review of Economic Studies*, 47:1–20, 1980.

[Richardson, 1996] T. Richardson. A discovery algorithm for directed cyclic graphs. In E. Horvitz and F. Jensen, editors, *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 454–461. Morgan Kaufmann, San Francisco, 1996.

[Robins and Greenland, 1992] J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155, 1992.

[Robins, 1986] J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period: Aoplications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[Robins, 1995] J.M. Robins. Discussion of Causal diagrams for empirical research by J. Pearl. *Biometrika*, 82:695–698, 1995.

[Rosenbaum and Rubin, 1983] P. Rosenbaum and D. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrica*, 70:41–55, 1983.

[Rubin, 1974] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[Rubin, 1990] D.B. Rubin. Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5:472–480, 1990.

[Shafer, 1996] G. Shafer. *The Art of Causal Conjecture.* MIT Press, Cambridge, MA, 1996.

[Shipley, 1997] B. Shipley. An inferential test for structural equations models based on directed acyclic graphs and its nonparametric equivalents. Technical report, Départment de Biologie, Université de Sherbrooke, Canada, 1997.

[Simon, 1953] H.A. Simon. Causal ordering and identifiability. In Wm. C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 49–74. Wiley and Sons, Inc., 1953.

[Sobel, 1990] M.E. Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55:495–515, 1990.

[Sobel, 1995] M.E. Sobel. Causal inference in the social and behavioral sciences. In G. Arminger, C. Clogg, and M.E. Sobel, editors, *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pages 1–37. Plenum Press, New York, 1995.

[Spirtes and Richardson, 1996] P. Spirtes and T. Richardson. A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, 1996.

[Spirtes and Verma, 1992] P. Spirtes and T. Verma. Equivalence of causal models with latent variables. Technical Report CMU-PHIL-33, Department of Philosophy, Carnegie Mellon University, Pittsburgh, 1992.

[Spirtes et al., 1993] P. Spirtes, C. Glymour, and R. Schienes. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.

[Spirtes et al., 1996] P. Spirtes, T. Richardson, C. Meek, R. Scheines, and C. Glymour. Using d-separation to calculate zero partial correlations in linear models with correlated errors. Technical Report CMU-PHIL-72, Department of Philosophy, Carnegie Mellon University, Pittsburgh, 1996.

[Spirtes et al., 1998] P. Spirtes, T. Richardson, C. Meek, R. Scheines, and C. Glymour. Using path diagrams as a structural equation modeling tool. *Socioligical Methods and Research*, 27(2), November 1998. This volume.

[Stelzl, 1986] I. Stelzl. Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, 21:309–331, 1986.

[Steyer et al., 1996] R. Steyer, S. Gabler, and A.A. Rucai. Individual causal effects, average causal effects, and unfoundedness in regression models. In F. Faulbaum and W. Bandilla, editors, *SoftStat'95, Advances in Statistical Software 5*, pages 203–210. Lucius & Lucius, Stuttgart, 1996.

[Strotz and Wold, 1960] R.H. Strotz and H.O.A. Wold. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica*, 28:417–427, 1960.

[Verma and Pearl, 1988] T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*, pages 352–359. Mountain View, CA, 1988. Also in R. Schachter, T.S. Levitt, and L.N. Kanal (Eds.), *Uncertainty in AI 4*, Elesevier Science, 69–76, 1990.

[Verma and Pearl, 1990] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In P.P. Bonissone et al, editor, *Uncertainty in Artificial Intelligence 6*, pages 220–227. Elsevier Science, Cambridge, MA, 1990.

[Wainer, 1991] H. Wainer. Adjusting for differential base-rates: Lord's paradox again. *Psychological Bulletin*, 109:147–151, 1991.

[Weinberg, 1993] C.R. Weinberg. Toward a clearer definition of confounding. *American Journal of Epidemiology*, 137:1–8, 1993.

[Wermuth, 1992] N. Wermuth. On block-recursive regression equations. *Brazilian Journal of Probability and Statistics*, 6:1–56, 1992. (with discussion).

[Wright, 1921] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

[Wright, 1923] S. Wright. The theory of path coefficients: A reply to Niles' criticism. *Genetics*, 8:239–255, 1923.