A causal calculus for statistical research, with applications to observational and experimental studies

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

Abstract

Many statisticians are reluctant to deal with problems involving causal considerations because we lack the mathematical notation for distinguishing causal influence from statistical association. To address this problem, a notation is proposed that admits two conditioning operators: ordinary Bayes conditioning, $P(y|X = x)$, and causal conditioning, $P(y|set(X = x))$, that is, conditioning $P(y)$ on holding X constant (at x) by external intervention. This distinction, which will be supported by three rules of inference, will permit us to derive probability expressions for the combined effect of observations and interventions.

The resulting calculus yields simple solutions to a number of interesting problems in causal inference and should allow rank-and-file researchers to tackle practical problems that are generally considered too hard, or impossible. Examples are:

1. Deciding whether the information available in a given observational study is sufficient for obtaining consistent estimates of causal effects.
2. Deriving algebraic expressions for causal effect estimands.
3. Selecting measurements that would render randomized experiments unnecessary.
4. Selecting a set of indirect (randomized) experiments to replace direct experiments that are either infeasible or too expensive.
5. Predicting (or bounding) the efficacy of treatments from randomized trials with imperfect compliance.

Starting with nonparametric specification of structural equations, the paper establishes the semantics necessary for a theory of interventions, presents the three rules of inference, demonstrates the use of the resulting calculus on a number of examples, and establishes an operational definition of structural equations.

Key words: Causal inference, graph models, treatment effect, structural equations.

1 Introduction

The calculus introduced in this paper is aimed at helping researchers communicate qualitative assumptions about cause-effect relationships, elucidate the ramifications of such assumptions, and derive causal inferences from a combination of assumptions, experiments, and data.

The basic philosophy of the proposed method can best be illustrated through the following example [Cochran 1957]. Consider an experiment in which soil fumigants (X) are used to increase oat crop yields (Y) by controlling the eelworm population (Z) but may also have direct effects (both beneficial and adverse) on yields beside the control of eelworms. We wish to assess the total effect of the fumigants on yields when this classical experimental setup is complicated by several factors. First, controlled randomized experiments are infeasible – farmers insist on deciding for themselves which plots are to be fumigated. Second, we suspect that farmers' choice of treatment is predicated on last year's eelworm population (Z_0), an unknown quantity, and that last year's eelworm population is strongly correlated with this year's population — thus we have a classical case of confounding bias, which interferes with the assessment of treatment effects, regardless of sample size. Fortunately, through laboratory analysis of soil samples, we can determine the eelworm populations before and after the treatment and, furthermore, because the fumigants are known to be active for a short period only, we can safely assume that they do not affect the growth of eelworms surviving the treatment. However, the survival of eelworms past the application of the fumigants depends on the population of birds (and other predators) which is correlated, in turn, with last year's eelworm population and hence with the treatment itself.

The method proposed in this paper permits the investigator to translate complex considerations of this sort into a formal language, thus facilitating the following tasks:

1. Explicate the assumptions underlying the model.

2. Decide whether the assumptions are sufficient for obtaining consistent estimates of the target quantity: the total effect of the fumigants on yields.

3. If the answer to item 2 is affirmative, the method provides a closed-form expression for the target quantity, in terms of distributions of observed quantities.

4. If the answer to item 2 is negative, the method suggests a set of observations and experiments which, if performed, would render a consistent estimate feasible.

The first step in this analysis is to construct a causal diagram such as the one given in Figure 1 which represents the investigator's understanding of the major causal influences among measurable quantities in the domain. For example, the quantities Z_1, Z_2, and Z_3 represent, respectively, the eelworm population (both size and type) before treatment, after treatment, and at the end of the season. Z_0 represents last year's eelworm population; because it is an unknown quantity, it is denoted by a hollow circle, as is the quantity B, the population of birds and other predators. Links in the diagram are of two kinds: those that connect unmeasured quantities are designated by dashed arrows, those connecting measured quantities by solid arrows. The substantive assumptions embodied in the diagram are negative causal assertions which are conveyed through the links *missing* from the diagram. For example, the missing arrow between Z_1 and Y signifies the investigator's understanding

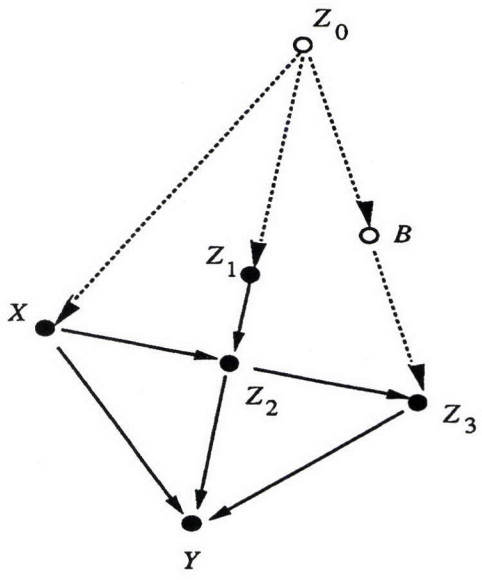

Figure 1:
A causal diagram representing the effect of fumigants (X) *on yields* (Y).

that pre-treatment eelworms can not affect oat plats directly; their entire influence on oat yields is mediated by post-treatment conditions, namely Z_2 and Z_3.

The proposed method allows an investigator to inspect the diagram of Figure 1 and conclude immediately that:

1. The total effect of X on Y can be estimated consistently from the observed distribution of X, Z_1, Z_2, Z_3, and Y.

2. The total effect of X on Y (assuming discrete variables) is given by the formula

$$P(y|\hat{x}) = \sum_{z_1}\sum_{z_2}\sum_{z_3} P(y|z_2,\ z_3,\ x)P(z_2|z_1,\ x)\sum_{x'} P(z_3|z_1,\ z_2,\ x')P(z_1,\ x') \quad (1)$$

   where $P(y|\hat{x})$ stands for the probability of achieving a yield level of $Y = y$ given that the treatment is *set* to level $X = x$ by external intervention.

3. A consistent estimation of the total effect of X on Y would not be feasible if Y were confounded with Z_3; however, confounding Z_2 and Y will not invalidate the formula for $P(y|\hat{x})$.

These conclusions can be obtained either by analyzing the graphical properties of the diagram or by performing a sequence of symbolic derivations, governed by the diagram, which gives rise to causal effect formulas such as Eq. (1). This paper establishes a calculus that systematizes these derivations.

2 Graphical Models and the Manipulative Account of Causation

The usefulness of directed acyclic graphs (DAGs) as economical schemes for representing conditional independence assumptions is well acknowledged in the literature [Pearl 1988]. This usefulness stems from the existence of a graphical criterion, called d-separation, which identifies each and every conditional independency that is implied by the product decomposition

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid \mathbf{pa}_i) \tag{2}$$

where $\mathbf{pa}_i$ are realizations of the variables corresponding to the direct predecessors (called *parents*) of X_i in a DAG G.

The use of DAGs as carriers of independence assumptions has also been instrumental in predicting the effect of interventions when DAGs are given a causal interpretation [Spirtes et al. 1993, Pearl 1993]. In [Pearl 1993], for example, interventions were treated as variables in an augmented probability space, and their effects were obtained by ordinary conditioning.

In this paper we will pursue a different (though equivalent) causal interpretation of DAGs, based on nonparametric structural equations, which owes its roots to early works in econometrics [Frisch 1938, Haavelmo 1943, Simon 1953]. In this account, assertions about causal influences, such as those specified by the links in Figure 1, stand for *autonomous* physical mechanisms among the corresponding quantities, and these mechanisms can be represented as functional relationships perturbed by random disturbances. In other words, each child-parent family in a DAG G represents a deterministic function

$$X_i = f_i(\mathbf{pa}_i, \epsilon_i), \qquad i = 1, ..., n \tag{3}$$

where $\mathbf{pa}_i$ are the parents of variable X_i in G, and ϵ_i, $0 < i \leq n$, are mutually independent, arbitrarily distributed random disturbances [Pearl & Verma 1991]. These disturbance terms represent independent exogenous factors that the investigator chooses not to include in the analysis. If any of these factors is judged to be influencing two or more variables (thus violating the independence assumption), then that factor must enter the analysis as an unmeasured (or latent) variable, to be represented in the graph by a hollow node, such as Z_0 and B in Figure 1. For example, the causal assumptions conveyed by the model in Figure 1 correspond to the following set of equations:

$$\begin{array}{ll} Z_0 = f_0(\epsilon_0) & Z_2 = f_2(X, Z_1, \epsilon_2) \\ B = f_B(Z_0, \epsilon_B) & Z_3 = f_3(B, Z_2, \epsilon_3) \\ Z_1 = f_1(Z_0, \epsilon_1) & Y = f_Y(X, Z_2, Z_3, \epsilon_Y) \\ X = f_X(Z_0, \epsilon_X) & \end{array} \tag{4}$$

The equational model in (3) is the nonparametric analogue of the so-called structural equations model in econometrics [Goldberger 1973], with one exception: the functional form of the equations as well as the distribution of the disturbance terms will remain unspecified. In contrast to ordinary algebraic equations, the equality signs in structural equations convey the asymmetrical relation of "is determined by", and should more accurately be

represented by an assignment symbols, as in computer programs.[1] Because of this asymmetry, structural equations communicate stable *counterfactual* information, thus forming a clear correspondence between causal diagrams and Rubin's model of potential response [Rubin 1974, Holland 88]. For example, the equation for Y states that regardless of what we currently observe about Y, and regardless of any changes that might occur in other equations, if $(X, Z_2, Z_3, \epsilon_Y)$ were to assume the values $(x, z_2, z_3, \epsilon_Y)$, respectively, Y would take on the value dictated by the function f_Y. Thus, the corresponding potential response variable in Rubin's model $Y_{(x)}$ (read: the value that Y would take if X were x) becomes a deterministic function of Z_2, Z_3 and ϵ_Y and can be considered a random variable whose distribution is determined by those of Z_2, Z_3 and ϵ_Y.

Characterizing each child-parent relationship as a deterministic function, instead of the usual conditional probability $P(x_i \mid \mathbf{pa}_i)$, imposes equivalent independence constraints on the resulting distributions and leads to the same recursive decomposition that characterizes DAG models (see Eq. (2)). This occurs because each ϵ_i is independent on all nondescendants of X_i. However, the functional characterization $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ also provides a convenient languages for specifying how the resulting distribution would change in response to external interventions. This is accomplished by encoding each intervention as an alteration on a select subset of functions, while keeping the others intact. Once we know the identity of the mechanisms altered by the intervention and the nature of the alteration, the overall effect of the intervention can be predicted by modifying the corresponding equations in the model and using the modified model to compute a new probability function.

The simplest type of external intervention is one in which a single variable, say X_i, is forced to take on some fixed value x_i. Such an intervention, which we call *atomic*, amounts to lifting X_i from the influence of the old functional mechanism $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ and placing it under the influence of a new mechanism that sets the value x_i while keeping all other mechanisms unperturbed. Formally, this atomic intervention, which we denote by $set(X_i = x_i)$, or $set(x_i)$ for short, amounts to removing the equation $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ from the model and substituting $X_i = x_i$ in the remaining equations. The new model thus created represents the system's behavior under the intervention $set(X_i = x_i)$ and, when solved for the distribution of X_j, yields the causal effect of X_i on X_j, denoted $P(x_j|\hat{x}_i)$.[2] More generally, when an intervention forces a subset X of variables to attain fixed values x, then a subset of equations is to be pruned from the model given in Eq. (3), one for each member of X, thus defining a new distribution over the remaining variables, which completely characterizes the effect of the intervention. We therefore define:

Definition 2.1 (*causal effect*) *Given two disjoint sets of variables, X and Y, the* causal effect *of X on Y is a function from X to the space of probability distributions on Y. For each realization x of X, $P(y|\hat{x})$ gives the probability of $Y = y$ induced by deleting from the model (8) all equations corresponding to variables in X and substituting $X = x$ in the remaining equations.* □

[1] I have found that economists, by and large, are not aware of this distinction.

[2] An explicit translation of interventions to "wiping out" equations from the model was first proposed by [Strotz & Wold 1960] and later used in [Fisher 1970] and [Sobel 1990]. Graphical ramifications of this interpretation were explicated first in [Spirtes et al. 1993] and later in [Pearl 1993]. An equivalent mathematical model, using event trees has been introduced by [Robins 1986, pp. 1422-1425].

Clearly the graph corresponding to the reduced set of equations is an edge subgraph of G from which all arrows entering X have been pruned. We will denote this subgraph by $G_{\overline{X}}$.

Regardless of whether we represent interventions as a modification of an existing model or as part of an augmented model, the result is a well-defined transformation between the pre-intervention and the post-intervention distributions. In the case of an atomic intervention $set(X_i = x_i')$, this transformation can be expressed in a simple algebraic formula that follows immediately from Eq. (3) and Definition 2.1:[3]

$$P(x_1, ..., x_n|\hat{x}_i') = \begin{cases} \frac{P(x_1,...,x_n)}{P(x_i \mid \mathbf{pa}_i)} & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases} \tag{5}$$

This formula reflects the removal of the term $P(x_i \mid \mathbf{pa}_i)$ from the product decomposition of Eq. (2), since $\mathbf{pa}_i$ no longer influence X_i. Graphically, the removal of this term is equivalent to removing the links between $\mathbf{pa}_i$ and X_i while keeping the rest of the network intact. Clearly, an intervention $set(x_i)$ can affect only the descendants of X_i in G.

The immediate implication of Eq. (5) is that, given the structure of the causal diagram G in which all variables are observable, one can infer post-intervention distributions from pre-intervention distributions; hence, we can reliably estimate the effects of interventions from passive (i.e., nonexperimental) observations. Of course, Eq. (5) does not imply that we can always substitute observational studies for experimental studies, as this would require estimation of $P(x_i \mid \mathbf{pa}_i)$. The mere identification of $\mathbf{pa}_i$ (i.e., the direct causal factors of X_i) requires substantive causal knowledge of the domain which is often unavailable. Moreover, even when we have sufficient substantive knowledge to structure the causal diagram (as in Figure 1) and identify $\mathbf{pa}_i$, some members of $\mathbf{pa}_i$ may be unobservable, or *latent*, thus preventing estimation of $P(x_i|\mathbf{pa}_i)$. Fortunately, there are conditions for which a consistent estimate of $P(x_j|\hat{x}_i)$ can be obtained even when the $\mathbf{pa}_i$ variables are latent and, moreover, simple graphical tests can tell us when such conditions are satisfied.

3 Controlling Confounding Bias

3.1 The Back-Door Criterion

Assume we are given a causal diagram G together with nonexperimental data on a subset $\mathbf{X}_o$ of observed variables in G and we wish to estimate what effect the intervention $set(X_i = x_i)$ would have on some response variable X_j. In other words, we seek to estimate $P(x_j|\hat{x}_i)$ from a sample estimate of $P(\mathbf{X}_o)$.

The variables in X_0 are commonly known as concomitants [Cox 1958]. In experimental studies, concomitants are used to reduce errors due to uncontrolled variations from sample to sample. In observational studies, concomitants are used to reduce confounding bias due to spurious correlations between treatment and response. The condition that qualifies a set $\mathbf{S}$ of concomitants as sufficient for identifying causal effect has been given a variety

[3]Eq. (5) can also be obtained from the G-computation formula of [Robins 1986, p. 1423] and the Manipulation Theorem of [Spirtes et al. 1993]. According to this source, Eq. (5) was "independently conjectured by Fienberg in a seminar in 1991".

of formulations, all requiring conditional independence judgments involving counterfactual variables [Rosenbaum & Rubin 1983, Pratt & Schlaifer 1988]. In [Pearl 1993] it was shown that such judgments can be translated to a simple d-separation conditions in the diagram G, which we name the *back-door criterion*:

Definition 3.1 (*back-door*) *A set of variables* $\mathbf{S}$ *is said to satisfy the* back-door *criterion relative to an ordered pair of variables* (X_i, X_j) *in a DAG* G *if*

1. *no node in* $\mathbf{S}$ *is a descendant of* X_i, *and*

2. $\mathbf{S}$ *blocks every path between* X_i *and* X_j *which contains an arrow into* X_i.

Similarly, if X *and* Y *are two disjoint subsets of nodes in* G, *then* $\mathbf{S}$ *is said to satisfy the back-door criterion relative to* (X, Y) *if it satisfies the criterion relative to any pair* (x, y) *such that* $x \in X$ *and* $y \in Y$. □

The name back-door echoes condition 2, which requires that only paths with arrows pointing at X_i be d-separated; these paths can be viewed as entering X_i through the back door. In Figure 2, for example, the sets $\mathbf{S}_1 = \{X_3, X_4\}$ and $\mathbf{S}_2 = \{X_4, X_5\}$ meet the back-door criterion, but $\mathbf{S}_3 = \{X_4\}$ does not because X_4 does not block the path $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$. Thus, we have obtained a simple graphical criterion for selecting a set of covariates which, if observed, would enable the identification of causal effects from nonexperimental data. An equivalent, though more complicated, graphical criterion is given in Theorem 7.1 of [Spirtes et al. 1993]. We summarize this finding in a theorem, after formally defining "identifiability".

Definition 3.2 (*identifiability*) *The causal effect of* X *on* Y *is said to be* identifiable *if the quantity* $P(y|\hat{x})$ *can be computed uniquely from the joint distribution of the observed variables. Identifiability means that* $P(y|\hat{x})$ *can be estimated consistently from an arbitrarily large sample randomly drawn from the joint distribution.* □

Theorem 3.3 *If a set of variables* Z *satisfies the back-door criterion relative to* (X, Y) *and* $P(x, z) > 0$, *then the causal effect of* X *on* Y *is identifiable and is given by the formula*

$$P(y|\hat{x}) = \sum_z P(y|x, z)P(z) \tag{6}$$

□

Reducing Rubin's ignorability conditions to the graphical criterion of Definition 3.1 replaces judgments about counterfactual interactions with formal procedures that can be applied to causal diagrams of any size and shape. The reduction to a graphical criterion also facilitates the search for an optimal set of concomitants, namely, a set Z that minimizes measurement cost or sampling variability.

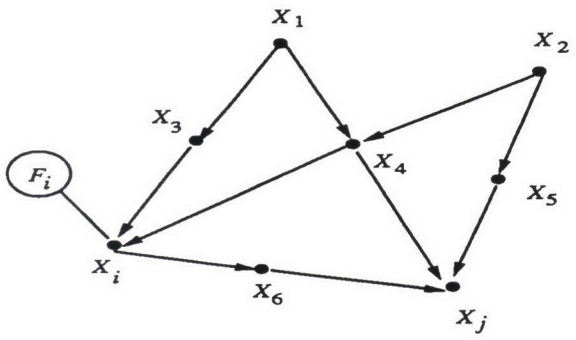

Figure 2:
A DAG representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ *(or* $\{X_4, X_5\}$*) yields an unbiased estimate of* $P(x_j|\hat{x}_i)$.

3.2 Other Graphical Criteria

The control of confounding bias does not end with the back-door estimand of Definition 3.1; an orthogonal estimand, worthy of the name "the front-door criterion", may complement the latter in cases where we cannot find observed covariates **S** satisfying the back-door conditions. Consider variable X_6 in Figure 2, and assume that it is the only observed variable in the graph, beside X_i and X_j. Clearly, X_6 does not satisfy any of the back-door conditions because (1) it is a descendant of X_i, and (2) it does not block any of the back-door paths between X_i and X_j. However, measurements of X_6 can nevertheless facilitate a consistent estimation of $P(x_j|\hat{x}_i)$. This can be shown either using the of the intervention calculus of Section 4 or by reducing the expression for $P(x_j|\hat{x}_i)$ to formulae computable from the observed distribution function $P(x_i, x_6, x_j)$ [Pearl 1994c]. To that end, let us denote by U the compound variable consisting of all confounding variables between X_i and X_j (i.e., $U = \{X_1, ..., X_5\}$ in Figure 2), and further denote X_i by X and X_j by Y. All together, we now have a structure depicted in Figure 3 below, containing one unobserved variable, U, three observed variables X, Z, Y, with Z mediating the interaction between X and Y. We will also assume that $P(x, z) > 0$ for all values of x and z.

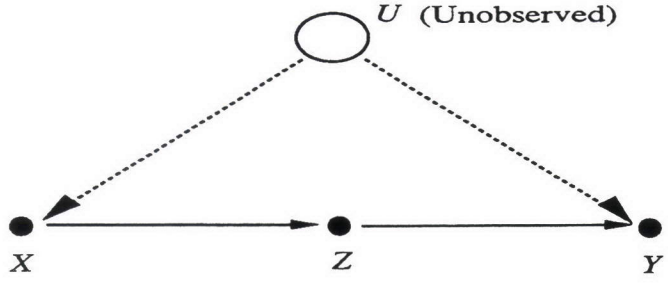

Figure 3:

From Eq. (4), the causal effect of X on Y is given by

$$P(y|\hat{x}) = \sum_u P(y|x, u)P(u) \tag{7}$$

Using the two conditional independence claims embodied in the graph of Figure 3, it is possible to eliminate u from the rhs of (7) and obtain:

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{8}$$

We summarize this result by a theorem,

Theorem 3.4 *If a variable Z satisfies the following conditions relative to an ordered pair of variables (X, Y),*

1. *Z intercepts all direct paths from X to Y*
2. *There is no back-door path between X and Z, nor between Z and Y.*
3. *The relation between X and Z is non-deterministic, i.e., $P(x, z) > 0$*

then the causal effect of X on Y is identifiable and is given by the formula in Eq. (8).

The graphical criterion of Theorem 3.4 uncovers many new structures that permit the identification of causal effects from nonexperimental observations. In contrast, most of the literature on statistical experimentation considers the measurement of intermediate variables, affected by the action, to be useless, if not harmful, for causal inference [Cox 1958, Pratt & Schlaifer 198 The relevance of such structures in practical situations can be seen, for instance, if we identify X with smoking, Y with lung cancer, Z with the amount of tar deposited in a subject's lungs, and U with an unobserved carcinogenic genotype that, according to the tobacco industry, also induces an inborn craving for nicotine. In this case, Eq. (8) would provide us with the means to quantify, from nonexperimental data, the causal effect of smoking on cancer. (Assuming, of course, that the data $P(x, y, z)$ is made available and that we believe that smoking does not have any direct causal effect on lung cancer except that mediated by tar deposits).

We should remark, though, that having obtained nonparametric estimands for causal effects does not imply that one should refrain from using parametric forms in the estimation phase of the study. Prior information about shapes of distributions and the nature of causal interactions can be incorporated into the analysis by limiting the distributions in the estimand formulas to specific parametric family of functions. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand given in Eq. (8) can be converted to the product

$$E(Y|\hat{x}) = R_{xz}\beta_{zy \cdot x}x \tag{9}$$

where $\beta_{zy \cdot x}$ is the standardized regression coefficient [Pearl 1994a], and the estimation problem reduces to that of estimating regression coefficients (e.g., by least-squares). More sophisticated estimation techniques, tailored specifically for causal inference, can be found in [Robins 1989, Sec. 17][Robins et al. 1992, pp. 331-333]. To handle more elaborate structures, including multiple Z variables, nested combinations of back-door and front-door patterns, and concurrent "set" operations, we now introduce a symbolic calculus of intervention.

4 A Calculus of Intervention

This section establishes a set of inference rules by which probabilistic sentences involving actions and observations can be transformed into other such sentences, thus providing a syntactic method of deriving (or verifying) claims about interventions. We will assume that we are given the structure of a causal diagram G in which some of the nodes are observable while the others remain unobserved. Our main problem will be to facilitate the syntactic derivation of causal effect expressions of the form $P(y|\hat{x})$, where X and Y stand for any subsets of observed variables. By derivation we mean step-wise reduction of the expression $P(y|\hat{x})$ to an equivalent expression involving standard probabilities of observed quantities. Whenever such reduction is feasible, the causal effect of X on Y is identifiable (see Definition 3.2).

4.1 Preliminary Notation

Let X, Y, and Z be arbitrary disjoint sets of nodes in a DAG G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$ (see Figure 4 for illustration). Finally, the expression $P(y|\hat{x}, z) \triangleq P(y, z|\hat{x})/P(z|\hat{x})$ stands for the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x.

4.2 Inference Rules

Armed with this notation we are now able to formulate the three basic inference rules of the proposed calculus. Proofs are provided in [Pearl 1994c].

Theorem 4.1 *Let G be a DAG associated with a causal model as defined in Eq. (3), and let P stand for the probability distribution of the variables in the models. For any disjoint subsets of variables X, Y, Z, and W we have:*

Rule 1 *Insertion/deletion of observations*

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{10}$$

Rule 2 *Action/observation exchange*

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{11}$$

Rule 3 *Insertion/deletion of actions*

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X,\ W)_{G_{\overline{X},\ \overline{Z(W)}}} \tag{12}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Each of the inference rules above follows from the basic interpretation of the "$\hat{x}$" operator as a replacement of the causal mechanism that connects X to its pre-action parents by a new mechanism $X = x$ introduced by the intervening force. The result is a submodel characterized by the subgraph $G_{\overline{X}}$ (named "manipulated graph" in [Spirtes et al. 1993]) which supports all three rules.

Rule 1 reaffirms d-separation as a valid test for conditional independence in the distribution resulting from the intervention $set(X = x)$, hence the graph $G_{\overline{X}}$. This rule follows from the fact that deleting equations from the system does not introduce any dependencies among the remaining disturbance terms (see Eq. (3)).

Rule 2 provides a condition for an external intervention $set(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The condition amounts to $\{X \cup W\}$ blocking all back-door paths from Z to Y (in $G_{\overline{X}}$), since $G_{\overline{X}\underline{Z}}$ retains all (and only) such paths.

Rule 3 provides conditions for introducing (or deleting) an external intervention $set(Z = z)$ without affecting the probability of $Y = y$. The validity of this rule stems, again, from simulating the intervention $set(Z = z)$ by the deletion of all equations corresponding to the variables in Z (hence the graph $G_{\overline{XZ}}$).

Corollary 4.2 *A causal effect q: $P(y_1, ..., y_k|\hat{x}_1, ..., \hat{x}_m)$ is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 4.1, which reduces q into a standard (i.e., hat-free) probability expression.* □

Whether the three rules above are sufficient for deriving all identifiable causal effects remains an open question. However, the task of finding a sequence of transformations (if such exists) for reducing an arbitrary causal effect expression can be systematized and executed by efficient algorithms [Galles 1994]. As the next subsection illustrates, symbolic derivations using the hat notation are much more convenient than algebraic derivations that aim at eliminating the latent variables from standard probability expressions (as in Section 3.2).

4.3 Symbolic Derivation of Causal Effects: An Example

We will now demonstrate how these inference rules can be used to derive causal effect estimands in the structure of Figure 3 above. We will see that this structure permits us to quantify the effect of every atomic intervention, using much simpler computations than those used in the derivation of the front-door formula (Section 3.2).

The applicability of the inference rules requires that the d-separation condition holds in various subgraphs of G; the structure of each subgraph varies with the expressions to be manipulated. Figure 4 displays the graphs that will be needed for the derivations that follow.

Task-1, compute $P(z|\hat{x})$

This task can be accomplished in one step, since G satisfies the applicability condition for

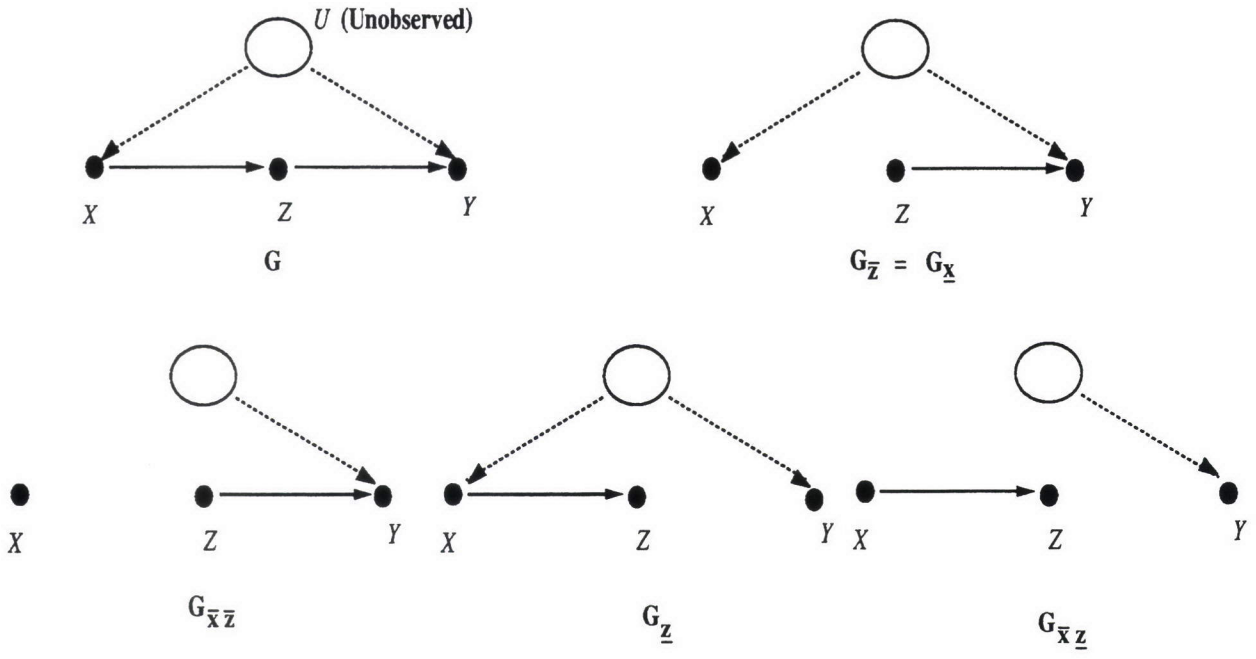

Figure 4:
Subgraphs of G used in the derivation of causal effects.

Rule 2; namely, $X \underline{\perp\!\!\!\perp} Z$ in $G_{\underline{X}}$ (because the path $X \leftarrow U \rightarrow Y \leftarrow Z$ is blocked by the collider at Y) and we can write

$$P(z|\hat{x}) = P(z|x) \tag{13}$$

Task-2, compute $P(y|\hat{z})$
Here we cannot apply Rule 2 to exchange $\hat{z}$ with z because $G_{\underline{Z}}$ contains a back-door path from Z to $Y : Z \leftarrow X \leftarrow U \rightarrow Y$. Naturally, we would like to block this path by conditioning on variables (such as X) that reside on that path. Symbolically, this involves conditioning and summing over all values of X,

$$P(y|\hat{z}) = \sum_x P(y|x,\hat{z})P(x|\hat{z}) \tag{14}$$

We now have to deal with two expressions involving $\hat{z}$, $P(y|x,\hat{z})$ and $P(x|\hat{z})$. The latter can be readily computed by applying Rule 3 for action deletion:

$$P(x|\hat{z}) = P(x) \text{ if } (Z \underline{\perp\!\!\!\perp} X)_{G_{\overline{Z}}} \tag{15}$$

noting that, indeed, X and Z are d-separated in $G_{\overline{Z}}$. (This can also be verified in G; manipulating Z will have no effect on X.) To reduce the former, $P(y|x,\hat{z})$, we consult Rule 2:

$$P(y|x,\hat{z}) = P(y|x,z) \text{ if } (Z \underline{\perp\!\!\!\perp} Y|X)_{G_{\underline{Z}}} \tag{16}$$

noting that X d-separates Z from Y in $G_{\underline{Z}}$. This allows us to write Eq. (14) as

$$P(y|\hat{z}) = \sum_x P(y|x,z)P(x) = E_x P(y|x,z) \tag{17}$$

which is a special case of the back-door formula (Eq. (6)). The legitimizing condition, $(Z \underline{\perp\!\!\!\perp} Y|X)_{G_{\underline{Z}}}$, offers yet another graphical test for the ignorability condition of [Rosenbaum & Rubin

Task-3, compute $P(y|\hat{x})$

Writing

$$P(y|\hat{x}) = \sum_z P(y|z,\hat{x})P(z|\hat{x}) \tag{18}$$

we see that the term $P(z|\hat{x})$ was reduced in Eq. (13) but that no rule can be applied to eliminate the "hat" symbol $\hat{}$ from the term $P(y|z,\hat{x})$. However, we can add a $\hat{}$ symbol to this term via Rule 2

$$P(y|z,\hat{x}) = P(y|\hat{z},\hat{x}) \tag{19}$$

since the applicability condition $(Y \underline{\perp\!\!\!\perp} Z|X)_{G_{\overline{X}\underline{Z}}}$, holds true (see Figure 4). We can now delete the action $\hat{x}$ from $P(y|\hat{z},\hat{x})$ using Rule 3, since $Y \underline{\perp\!\!\!\perp} X|Z$ holds in $G_{\overline{XZ}}$. Thus, we have

$$P(y|z,\hat{x}) = P(y|\hat{z}) \tag{20}$$

which was calculated in Eq. (17). Substituting Eqs. (17), (20), and (13) back into Eq. (18) finally yields

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x',z)P(x') \tag{21}$$

which is identical to the front-door formula of Eq. (8).

Task-4, compute $P(y,z|\hat{x})$

$$P(y,z|\hat{x}) = P(y|z,\hat{x})P(z|\hat{x})$$

The two terms on the r.h.s. were derived before in Eqs. (13) and (20), from which we obtain

$$\begin{aligned} P(y,z|\hat{x}) &= P(y|\hat{z})P(z|x) \\ &= P(z|x)\textstyle\sum_{x'} P(y|x',z)P(x') \end{aligned} \tag{22}$$

Task-5, compute $P(x,y|\hat{z})$

$$\begin{aligned} P(x,y|\hat{z}) &= P(y|x,\hat{z})P(x|\hat{z}) \\ &= P(y|x,z)P(x) \end{aligned} \tag{23}$$

The first term on the r.h.s. is obtained by Rule 2 (licensed by $G_{\underline{Z}}$) and the second term by Rule 3 (as in Eq. (15)).

Note that in all the derivations the graph G has provided both the license for applying the inference rules and the guidance for choosing the right rule to apply.

4.4 Causal Inference by Surrogate Experiments

Suppose we wish to learn the causal effect of X on Y when X and Y are confounded and, for practical reasons of cost or ethics, we cannot control X by randomized experiment, nor can we find observed covariates that, if adjusted for, would eliminate the confounding effect between X and Y. The question arises whether $P(y|\hat{x})$ can be identified by randomizing a *surrogate* variable Z, which is easier to control than X. For example, if we are interested in assessing the causal effect of cholesterol levels (X) on heart disease (Y), a reasonable experiment to conduct would be to control subjects' diet (Z), rather than exercising direct control over cholesterol levels in subjects' blood.

Formally, this problem amounts to transforming $P(y|\hat{x})$ into expressions in which only members of Z obtain the hat symbol. Using Theorem 4.1 it can be shown [Pearl 1994c] that the following conditions are sufficient for admitting a surrogate variable Z:

1. X intercepts all directed paths from Z to Y, and,
2. $P(y|\hat{x})$ is identifiable in $G_{\overline{Z}}$.

Translated to our cholesterol example, this condition requires that there be no direct effect of diet on heart conditions and no confounding effect between cholesterol levels and heart disease, unless we can measure an intermediate variable between the two.

5 Graphical Tests of Identifiability

In the example of Section 4.3, we were able to compute all expressions of the form $P(r|\hat{s})$ where R and S are subsets of observed variables. In general, this will not be the case. For example, there is no general way of computing $P(y|\hat{x})$ from the observed distribution whenever the causal model contains the bow-pattern shown in Figure 5, in which X and Y are connected by both a causal link and a confounding arc. A confounding arc represents the existence in the diagram of a back-door path that contains only unobserved variables and has no converging arrows. For example, the path X, Z_0, B, Z_3 in Figure 1 can be represented as a confounding arc between X and Z_3. A bow-pattern represents an equation $Y = f_Y(X, U, \epsilon_Y)$ where U is unobserved and dependent on X. Such an equation does not permit the identification of causal effects since any portion of the observed dependence between X and Y may always be attributed to spurious dependencies mediated by U.

The presence of a bow-pattern prevents the identification of $P(y|\hat{x})$ even when it is found in the context of a larger graph, as in Figure 5(b). This is in contrast to linear models, where the addition of an arc to a bow-pattern can render $P(y|\hat{x})$ identifiable. For example, if Y is related to X via a linear relation $Y = bX + U$, where U is a zero-mean disturbance possibly correlated with X, then $b = \frac{\partial}{\partial x}E(Y|\hat{x})$ is not identifiable. However, adding an arc $Z \rightarrow X$ to the structure (that is, finding a variable Z that is correlated with X but not with U) would facilitate the computation of $E(Y|\hat{x})$ via the instrumental-variable formula [Angrist et al. 1993]:

$$b \triangleq \frac{\partial}{\partial x}E(Y|\hat{x}) = \frac{E(Y|z)}{E(X|z)} = \frac{R_{yz}}{R_{xz}} \tag{24}$$

In nonparametric models, adding an instrumental variable Z to a bow-pattern (Figure 5(b)) does not permit the identification of $P(y|\hat{x})$. This is a familiar problem in the analysis of clinical trials in which treatment assignment (Z) is randomized (hence, no link enters Z), but compliance is imperfect. The confounding arc between X and Y in Figure 5(b) represents unmeasurable factors which influence both subjects' choice of treatment (X) and subjects' response to treatment (Y). In such trials, it is not possible to obtain an unbiased estimate of the treatment effect $P(y|\hat{x})$ without making additional assumptions on the nature of the interactions between compliance and response. One can calculate bounds on $P(y|\hat{x})$ [Robins 1989][Manski 1990, Sec. 1g] and the upper and lower bounds may even coincide for certain types of distributions $P(x, y, z)$ [Balke & Pearl 1993], but there is no way of computing $P(y|\hat{x})$ for *every* distribution $P(x, y, z)$.

A general feature of nonparametric models is that the addition of arcs to a causal diagram can impede, but never assist, the identification of causal effects. This is because such addition reduces the set of d-separation conditions carried by the diagram and, hence, if a causal effect derivation fails in the original diagram, it is bound to fail in the augmented diagram as well. Conversely, any causal effect derivation that succeeds in the augmented diagram (by a sequence of symbolic transformations, as in Corollary 4.2) would succeed in the original diagram.

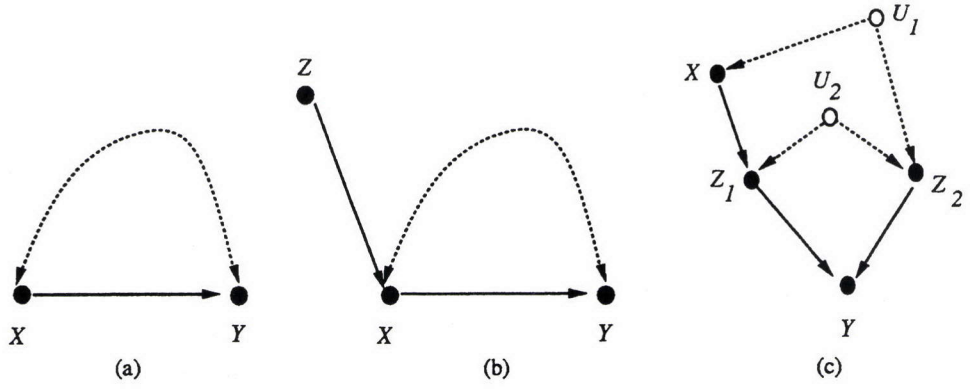

Figure 5:
(a) A bow-pattern: a confounding arc embracing a causal link $X \rightarrow Y$, thus preventing the identification of $P(y|\hat{x})$ even in the presence of an instrumental variable Z, as in (b). (c) A bow-less graph still prohibiting the identification of $P(y|\hat{x})$.

Our ability to compute $P(y|\hat{x})$ for pairs (x, y) of singleton variables does not ensure our ability to compute joint distributions, such as $P(y_1, y_2|\hat{x})$. Figure 5(c), for example, shows a causal diagram where both $P(z_1|\hat{x})$ and $P(z_2|\hat{x})$ are computable, but $P(z_1, z_2|\hat{x})$ is not. Consequently, we cannot compute $P(y|\hat{x})$. Interestingly, this diagram is the smallest graph that does not contain a bow-pattern and still presents an uncomputable causal effect.

Another interesting feature demonstrated by Figure 5(c) is that computing the effect of a joint action is often easier than computing the effects of its constituent singleton actions.[4]

[4]This was brought to my attention by James Robins, who has worked out many of these computations in the context of sequential treatment management. Eq. (25) for example, can be obtained from Robin's

Here, it is possible to compute $P(y|\hat{x}, \hat{z}_2)$ and $P(y|\hat{x}, \hat{z}_1)$, yet there is no way of computing $P(y|\hat{x})$. For example, the former can be evaluated by invoking Rule 2 in $G_{\overline{X}\underline{Z}_2}$, giving

$$P(y|\hat{x}, \hat{z}_2) = \sum_{z_1} P(y|z_1, \hat{x}, \hat{z}_2) P(z_1|\hat{x}, \hat{z}_2) = \sum_{z_1} P(y|z_1, x, z_2) P(z_1|x) \tag{25}$$

However, Rule 2 cannot be used to convert $P(z_1|\hat{x}, z_2)$ into $P(z_1|x, z_2)$ because, when conditioned on Z_2, X and Z_1 are d-connected in $G_{\underline{X}}$ (through the dashed lines). We conjecture, however, that whenever $P(y|\hat{x}_i)$ is computable for every singleton variable X_i, then $P(y|\hat{x}_1, \hat{x}_2, ..., \hat{x}_l)$ is computable as well, for any subset of variables $\{X_1, ...X_l\}$. In [Pearl 1994c], we provide a more complete road map for graphs that permits the identification of causal effects.

6 The Operational Meaning of Structural Equations

Traditionally, statisticians have approved of only one method of combining subject-matter considerations with statistical data: the Bayesian method of assigning subjective priors to distributional parameters. To incorporate causal information within the Bayesian framework, plain causal statements such as "Y is affected by X" must be converted into sentences capable of receiving probability values, e.g., counterfactuals. Indeed, this is how Rubin's model has achieved statistical legitimacy: causal judgments are expressed as constraints on probability functions involving counterfactual variables.

Causal diagrams offer an alternative language for combining data with causal information. This language simplifies the Bayesian route by accepting plain causal statements as its basic primitives. These statements, which merely identify whether a causal connection between two variables of interest exists, are commonly used in natural discourse and provide a natural way for scientists to communicate experience and organize knowledge. It is hoped, therefore, that the language of causal graphs will find applications in problems requiring substantial use of subject-matter considerations.

The language is not new. The use of diagrams and structural equations models to convey causal information has been quite popular in the social sciences and econometrics. Statisticians, however, have generally found these models suspect, perhaps because social scientists and econometricians have failed to provide an unambiguous definition of the empirical content of their models, that is, of the experimental conditions under which the outcomes are constrained by a given structural equation. As a result, even such basic notions as "structural coefficients" or "missing links" become the object of serious controversy [Freedman 1987] and conflicting interpretations [Wermuth 1992, Whittaker 1990, Cox & Wermuth 1993].

To a large extent, this history of controversy and miscommunication stems from the absence of an adequate mathematical notation for defining basic notions of causal modeling. Indeed, standard probabilistic notation cannot express the empirical content of the coefficient b in the structural equation $Y = bX + \epsilon_Y$ even if one is prepared to assume that ϵ_Y (an unobserved quantity) is uncorrelated with X. Nor can any probabilistic meaning be attached to the analyst's excluding from the equation certain variables that are highly correlated with X or Y.

G-computation algorithm [Robins 1986, p. 1423].

The notation developed in this paper gives these notions a clear empirical interpretation, because it permits one to specify precisely what is being held constant in a controlled experiment. The meaning of b is simply $\frac{\partial}{\partial x}E(Y|\hat{x})$, namely, the rate of change (in 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_Y and X are correlated and, moreover, the notion of randomization need not be invoked. Similarly, the analyst's decision as to which variables should be included in the equation for Y is based on a hypothetical controlled experiment in which several variables are controlled independently. A variable Z is excluded from the equation for Y if the analyst can identify some other variable (or a set of variables), say X, which, if held fixed, would prevent Z from influencing Y, that is, $P(y|\hat{x},\hat{z}) = P(y|\hat{x})$. In other words, variables that are excluded from the equation are not conditionally independent of Y given X, but rather conditionally independent of Y *setting* X. The operational meaning of the so called "disturbance term", ϵ_Y, is likewise demystified; ϵ_Y is defined by the difference between Y and the prediction of Y, based on *setting* all other observed variables.

The distinctions provided by the "hat" notation clarifies the empirical basis of structural equations and should make structural models more acceptable to statisticians. Moreover, since most scientific knowledge is organized around the operation of "holding X fixed," rather than "conditioning on X," the notation and calculus developed in this paper should provide a natural means for scientists to articulate subject-matter information, and to derive its logical consequences.

7 Extensions

Several extensions of the methods proposed in this paper are noteworthy. First, the analysis of atomic interventions can be generalized to complex policies in which a variable X is made to respond in a specified way to some set Z of other variables, say through a functional relationship $X = g(Z)$ or through a stochastic relationship whereby X is set to x with probability $P^*(x|z)$. In [Pearl 1994b] it is shown that computing the effect of such policies is equivalent to computing the expression $P(y|\hat{x},z)$.

A second extension concerns the use of the intervention calculus (Theorem 4.1) in nonrecursive models, that is, in causal diagrams involving directed cycles or feedback loops. The basic definition of causal effects in term of "wiping out" equations from the model (Definition 2.1) still carries over to nonrecursive systems [Strotz & Wold 1960, Sobel 1990], but then two issues must be addressed. First, the analysis of identification must ensure the stability of the remaining submodels [Fisher 1970]. Second, the d-separation criterion for DAGs must be extended to cover cyclic graphs as well. The validity of d-separation has been established for nonrecursive linear models and extended, using an augmented graph, to any arbitrary set of stable equations [Spirtes 1994]. However, the computation of causal effect estimands will be harder in cyclic networks, because symbolic reduction of $P(y|\hat{x})$ to hat-free expressions may require the solution of nonlinear equations.

Acknowledgment

Much of this investigation was inspired by [Spirtes et al. 1993], in which a graphical account of manipulations was first proposed. Phil Dawid, David Freedman, James Robins and Donald Rubin have provided genuine encouragement and valuable advice. The investigation also benefitted from discussions with Joshua Angrist, Peter Bentler, David Cox, David Galles, Arthur Goldberger, David Hendry, Paul Holland, Guido Imbens, Ed Leamer, Rod McDonald, John Pratt, Paul Rosenbaum, Keunkwan Ryu, Glenn Shafer, Michael Sobel, and David Tritchler. The research was partially supported by Air Force grant #F49620-94-1-0173, NSF grant #IRI-9200918, and Northrop-Rockwell Micro grant #93-124.

References

[Angrist et al. 1993] Angrist, J.D., Imbens, G.W., and Rubin, D.B., "Identification of causal effects using instrumental variables," Department of Economics, Harvard University, Cambridge, MA, Technical Report No. 136, June 1993.

[Balke & Pearl 1993] Balke, A., and Pearl, J., "Nonparametric bounds on treatment effects in partial compliance studies," UCLA Computer Science Department, Technical Report R-199, 1993. Short version in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, CA, 46-54, 1994.

[Cochran 1957] Cochran, W.G., "Analysis of covariance: Its nature and uses," *Biometrics*, 13, 261-281, 1957.

[Cox 1958] Cox, D.R., *The Planning of Experiments*, John Wiley and Sons, New York, 1958.

[Cox & Wermuth 1993] Cox, D.R., and Wermuth, N., "Linear dependencies represented by chain graphs," *Statistical Science*, 8(3), 204-218, 1993.

[Fisher 1970] Fisher, F.M., "A correspondence principle for simultaneous equation models," *Econometrica*, 38, 73-92, 1970.

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

[Frisch 1938] Frisch, R., "Statistical versus theoretical relations in economic macrodynamics," League of Nations Memorandum, 1938. (Reproduced in *Autonomy of Economic Relations*, Universitetets Socialokonomiske Institutt, Olso, 1948).

[Galles 1994] Galles, D., "Testing identifiability of causal effects," UCLA Computer Science Department, Technical Report R-225, in preparation.

[Goldberger 1973] Goldberger, A.S., *Structural Equation Models in the Social Sciences*, Seminar Press, New York, 1973.

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

[Holland 88] Holland, P.W., "Causal inference, path analysis, and recursive structural equations models," in C. Clogg (Ed.), *Sociological Methodology*, American Sociological Association, Washington, DC, pp. 449-484, 1988.

[Manski 1990] Manski, C.F., "Nonparametric bounds on treatment effects," *American Economic Review, Papers and Proceedings,* 80, 319-323, 1990.

[Pearl 1988] Pearl, J., *Probabilistic Reasoning in Intelligence Systems*, Morgan Kaufmann, San Mateo, CA, 1988. Revised 2nd printing, 1992.

[Pearl 1993] Pearl, J., "Comment: Graphical models, causality, and intervention," *Statistical Science*, 8(3), 266-269, 1993.

[Pearl 1994a] Pearl, J., "A note on testing exogeneity of instrumental variables," UCLA Computer Science Department, Technical Report R-211-S, 1994.

[Pearl 1994b] Pearl, J., "A Probabilistic Calculus of Actions," in *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, CA, 454-462, 1994.

[Pearl 1994c] Pearl, J., "Causal diagrams for experimental research," UCLA Computer Science Department, Technical Report (R-218-L), May 1994. To appear in *Biometrika.*

[Pearl & Verma 1991] Pearl, J. and Verma, T., "A theory of inferred causation," in J.A. Allen, R. Fikes and E. Sandewall (Eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, Morgan Kaufmann, San Mateo, CA, 441-452, 1991.

[Pratt & Schlaifer 1988] Pratt, J.W., and Schlaifer, R., "On the interpretation and observation of laws," *Journal of Econometrics*, 39, 23-52, 1990.

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

[Robins 1989] Robins, J.M., "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies," in L. Sechrest, H. Freeman, and A. Mulley (Eds.), *Health Service Research Methodology: A Focus on AIDS*, NCHSR, U.S. Public Health Service, pp. 113-159, 1989.

[Robins et al. 1992] Robins, J.M., Blevins, D., Ritter, G., and Wulfsohn, M., "*G*-Estimation of the Effect of Prophylaxis Therapy for Pneumocystis carinii Pneumonia on the Survival of AIDS Patients," *Epidemiology*, Vol. 3, Number 4, 319-336, 1992.

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

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

[Simon 1953] Simon, H.A., "Causal Ordering and Identifiability," in W.C. Hood and T.C. Koopmans (Eds.), *Studies in Econometric Method*, New York, NY, Chapter 3, 1953.

[Sobel 1990] Sobel, M.E., "Effect analysis and causation in linear structural equation models," *Psychometrika*, 55(3), 495-515, 1990.

[Spirtes 1994] Spirtes, P., "Conditional independence in directed cyclic graphical models for feedback," Department of Philosophy, Carnegie-Mellon University, Pittsburg, PA, Technical Report CMU-PHIL-53, May 1994.

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

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

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

[Whittaker 1990] Whittaker, J., *Graphical Models in Applied Multivariate Statistics*, John Wiley and Sons, Chichester, England, 1990.