Causal diagrams for empirical research

JUDEA PEARL
Cognitive Systems Laboratory, Computer Science Department,
University of California, Los Angeles, CA 90024, U.S.A.

SUMMARY

The primary aim of this paper is to show how graphical models can be used as a mathematical language for integrating statistical and subject-matter information. In particular, the paper develops a principled, nonparametric framework for causal inference, in which diagrams are queried to determine if the assumptions available are sufficient for identifying causal effects from nonexperimental data. If so the diagrams can be queried to produce mathematical expressions for causal effects in terms of observed distributions; otherwise, the diagrams can be queried to suggest additional observations or auxiliary experiments from which the desired inferences can be obtained.
Key words: Causal inference, graph models, structural equations, treatment effect.

1 INTRODUCTION

The tools introduced in this paper are 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 classical example due to Cochran (Wainer, 1989). 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 typical study is complicated by several factors. First, controlled randomized experiments are infeasible: farmers insist on deciding for themselves which plots are to be fumigated. Secondly, farmers' choice of treatment depends on last year's eelworm population (Z_0), an unknown quantity which 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. Insead, eelworms' growth

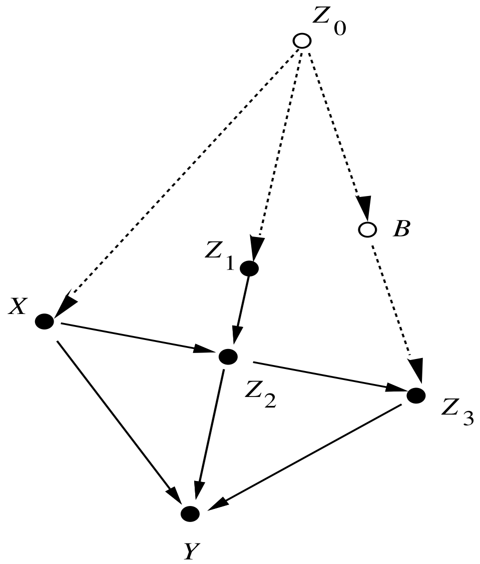

Figure 1: A causal diagram representing the effect of fumigants (X) on yields (Y).

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:

(i) Explicate the assumptions underlying the model.

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

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

(iv) If the answer to item (ii) 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 that pretreatment eelworms can not affect oat plants directly; their entire influence on oat yields is mediated by post-treatment conditions, namely Z_2 and Z_3. The purpose of the paper is not to validate or repudiate such domain-specific assumptions but, rather, to test whether a given set of assumptions is sufficient for quantifying causal effects from nonexperimental data, for example, estimating the total effect of fumigants on yields.

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

(i) 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.

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

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

where $pr(y|\check{x})$ 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.

(iii) 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 $pr(y|\check{x})$.

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).

The formal semantics of the causal diagrams used in this paper will be defined in Section 2, following a review of directed acyclic graphs as a language for communicating conditional independence assumptions. Subsection 2.2 introduces a causal interpretation of directed graphs based on nonparametric structural equations and demonstrates their use in predicting the effect of interventions. Section 3 will demonstrate the use of causal diagrams to control confounding bias in observational studies. We will establish two graphical conditions ensuring that causal effects can be estimated consistently from nonexperimental data. The first condition, named the back-door criterion, is equivalent to the ignorability condition of Rosenbaum & Rubin (1983). The second condition, named the front-door criterion, involves covariates that are affected by the treatment, and thus introduces new opportunities for causal inference. In Section 4, we introduce a symbolic calculus that permits the stepwise derivation of causal effect formulas of the type shown in (1). Using this calculus, Section 5 characterizes the class of graphs that permit the quantification of causal effects from nonexperimental data or from surrogate experimental designs.

2 GRAPHICAL MODELS AND THE MANIPULATIVE ACCOUNT OF CAUSATION

2.1 *Graphs and conditional independence*

The usefulness of directed acyclic graphs as economical schemes for representing conditional independence assumptions is well acknowledged in the literature (Pearl, 1988; Whittaker, 1990). This usefulness stems from the existence of graphical methods for identifying the conditional independence relationships that are implied by recursive product decompositions

$$pr(x_1, ..., x_n) = \prod_i pr(x_i \mid pa_i) \tag{2}$$

where pa_i stands for the realization of some subset of the variables that precede X_i in the order $(X_1, X_2, \ldots, X_n)$. If we construct a DAG in which the variables corresponding to pa_i are represented as the parents of X_i, also called adjacent predecessors or direct influences of X_i, then the independencies implied by the decomposition (2) can be read off the graph using the following test:

DEFINITION 1. (*d*-separation) Let X, Y, and Z be three disjoint subsets of nodes in a DAG G, and let p be any path between a node in X and a node in Y, where by a path we mean any succession of arcs, regardless of their directions. Then Z is said to block p if there is a node w on p satisfying one of the following two conditions: (*i*) w has converging arrows (along p) and neither w nor any of its descendants are in Z, or, (*ii*) w does not have converging arrows (along p) and w is in Z. Further, Z is said to d-separate X from Y, in G, written $(X \perp\!\!\!\perp Y|Z)_G$, iff Z blocks every path from a node in X to a node in Y. □

It can be shown that there is a one-to-one correspondence between the set of conditional independencies, $X \perp\!\!\!\perp Y|Z$ (Dawid, 1979), implied by the recursive decomposition of Eq. (2) and the set of triples (X, Z, Y) that satisfy the *d*-separation criterion in G (Geiger et al., 1990).

An alternative test for *d*-separation has been devised by Lauritzen et al. (1990), based on the notion of ancestral graphs. To test for $(X \perp\!\!\!\perp Y|Z)_G$, delete from G all nodes except those in $\{X, Y, Z\}$ and their ancestors, connect by an edge every pair of nodes that share a common child, and remove all arrows from the arcs. $(X \perp\!\!\!\perp Y|Z)_G$ holds iff Z is a cutset of the resulting undirected graph, separating nodes of X from those of Y. Additional properties of directed acyclic graphs and their applications to evidential reasoning in expert systems are discussed in Pearl (1988), Lauritzen & Spiegelhalter (1988), Spiegelhalter et al. (1993), and Pearl (1993a).

2.2 *Graphs as models of interventions*

The use of directed acyclic graphs as carriers of independence assumptions has also been instrumental in predicting the effect of interventions when these graphs are given a causal interpretation (Spirtes et al., 1993; Pearl, 1993b). In Pearl (1993b), 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 directed graphs, 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 are represented as functional relationships perturbed by random disturbances. In other words, each child-parent family in a directed graph G represents a deterministic function

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

where pa_i are the parents of variable X_i in G, and ϵ_i, $1 \leq i \leq n$, 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:

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

The equational model in (3) is the nonparametric analogue of the so-called structural equations model (Wright, 1921; Goldberger, 1973), with one exception: the functional form of the equations as well as the distribution of the disturbance terms will remain unspecified. The equality signs in structural equations convey the asymmetrical counterfactual relation of "is determined by", thus forming a clear correspondence between causal diagrams and Rubin's model of potential outcome (Rubin, 1974; Holland, 1988; Pratt & Schlaifer, 1988; Rubin, 1990). 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 $Y_{(x)}$, 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. The relation between graphical and counterfactual models is further analyzed in Pearl (1994a).

Characterizing each child-parent relationship as a deterministic function, instead of the usual conditional probability $pr(x_i \mid pa_i)$, imposes equivalent independence constraints on the resulting distributions and leads to the same recursive decomposition (2) that characterizes directed acyclic graph models. This occurs because each ϵ_i is independent on all nondescendants of X_i. However, the functional characterization $X_i = f_i(pa_i, \epsilon_i)$ also provides a convenient language 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 $X_i = f_i(pa_i, \epsilon_i)$ 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 $X_i = f_i(pa_i, \epsilon_i)$ 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 $pr(x_j|\check{x}_i)$. 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 (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. (causal effect) Given two disjoint sets of variables, X and Y, the causal effect of X on Y, denoted $pr(y|\check{x})$, is a function from X to the space of probability distributions on Y. For each realization x of $X, pr(y|\check{x})$ 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. □

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 translation were explicated first in Spirtes et al. (1993) and later in Pearl (1993b). A related mathematical model, using event trees has been introduced by Robins (1986, pp. 1422–1425).

Regardless of whether we represent interventions as a modification of an existing model as in Definition 2, or as a conditionalization in an augmented model (Pearl, 1993b), 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 (3) and Definition 2:

$$pr(x_1, ..., x_n|\check{x}_i') = \begin{cases} \frac{pr(x_1,...,x_n)}{pr(x_i \mid pa_i)} & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases} \quad . \tag{5}$$

This formula reflects the removal of the term $pr(x_i \mid pa_i)$ from the product in (2), since pa_i no longer influence X_i. Graphically, the removal of this term is equivalent to removing the links between pa_i and X_i while keeping the rest of the network intact. Equation (5) can also be obtained from the G-computation formula of Robins (1986, p. 1423) and the Manipulation Theorem of Spirtes et al. (1993) who state that this formula was "independently conjectured by Fienberg in a seminar in 1991". Clearly, an intervention $set(x_i)$ can affect only the descendants of X_i in G. Additional properties of the transformation defined in (5) are given in Pearl (1993b).

The immediate implication of (5) is that, given a causal diagram in which all parents of manipulated variables are observable, one can infer post-intervention distributions from pre-intervention distributions; hence, under such assumptions we can estimate the effects of interventions from passive i.e., nonexperimental observations. The aim of this paper, however, is to derive causal effects in situations such as Figure 1, where some members of pa_i may be unobservable, thus preventing estimation of $pr(x_i|pa_i)$. The next two sections provide simple graphical tests for deciding when $pr(x_j|\check{x}_i)$ is estimable in a given model.

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 V_0 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 $pr(x_j|\check{x}_i)$ from a sample estimate of $pr(V_0)$.

The variables in $V_0 \backslash \{X_i, X_j\}$, are commonly known as concomitants (Cox, 1958, p. 48). In observational studies, concomitants are used to reduce confounding bias due to spurious correlations between treatment and response. The condition that renders a set Z of concomitants sufficient for identifying causal effect, also known as ignorability, has been given a variety of formulations, all requiring conditional independence judgments involving counterfactual variables (Rosenbaum & Rubin, 1983; Pratt & Schlaifer, 1988). In Pearl (1993b) it is shown that such judgments are equivalent to a simple graphical test, named the "back-door criterion", which can be applied directly to the causal diagram.

DEFINITION 3. (back-door) A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a directed acyclic graph G if (i) no node in Z is a descendant of X_i, and (ii) Z 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 Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$. $\Box$

The name back-door echoes condition 2, which requires that only paths with arrows pointing at X_i be d-blocked; these paths can be viewed as entering X_i through the back door. In Figure 2, for example, the sets $Z_1 = \{X_3, X_4\}$ and $Z_2 = \{X_4, X_5\}$

meet the back-door criterion, but $Z_3 = \{X_4\}$ does not because X_4 does not block the path $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$. An equivalent, though more complicated, graphical criterion is given in Theorem 7.1 of Spirtes et al. (1993). An alternative criterion, using a single d-separation test will be established in Section 4.

We summarize this finding in a theorem, after formally defining "identifiability".

DEFINITION 4. (identifiability) The causal effect of X on Y is said to be identifiable if the quantity $pr(y|\check{x})$ can be computed uniquely from any positive distribution of the observed variables. □

Identifiability means that $pr(y|\check{x})$ can be estimated consistently from an arbitrarily large sample randomly drawn from the joint distribution. Note that, to prove non-identifiability, it is sufficient to present two sets of structural equations that induce identical distributions over observed variables but different causal effects.

THEOREM 1. If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula

$$pr(y|\check{x}) = \sum_z pr(y|x, z)pr(z) \tag{6}$$

□

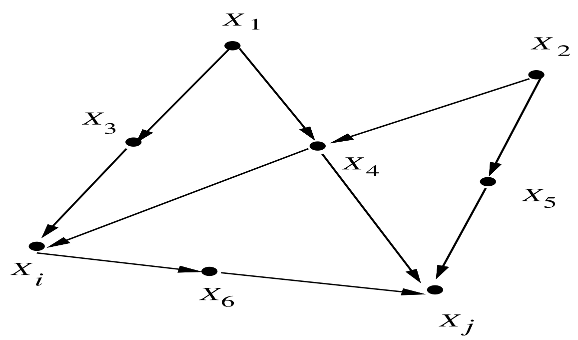

Figure 2: A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $pr(x_j|\check{x}_i)$.

Equation (6) represents the standard adjustment for concomitants Z when X is conditionally ignorable given Z (Rosenbaum & Rubin, 1983). Reducing ignorability conditions to the graphical criterion of Definition 3 replaces judgments about counterfactual dependencies with systematic procedures that can be applied to causal diagrams of any size and shape. The graphical criterion also enables the analyst to search for an optimal set of concomitants, namely, a set Z that minimizes measurement cost or sampling variability.

3.2 *The front-door criteria*

An alternative criterion, "the front-door criterion", may be used in cases where we cannot find observed covariates Z satisfying the back-door conditions. Consider the diagram in Figure 3, which obtains from Figure 2 in case variables $X_1, \ldots, X_5$ are unobserved. Although Z does not satisfy any of the back-door conditions, measurements of Z can nevertheless enable consistent estimation of $pr(y|\check{x})$. This can be shown by reducing the expression for $pr(y|\check{x})$ to formulae computable from the observed distribution function $pr(x, y, z)$.

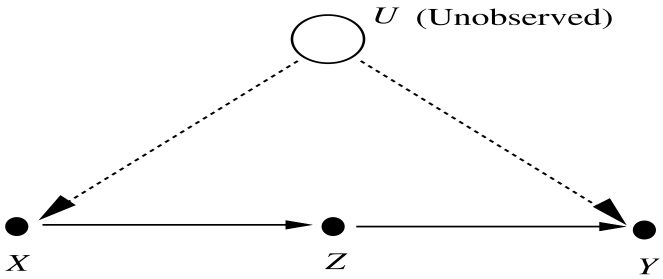

Figure 3: A diagram representing the front-door criterion.

The joint distribution associated with Figure 3 can be decomposed into

$$pr(x, y, z, u) = pr(u)pr(x|u)pr(z|x)pr(y|z, u) \tag{7}$$

and, from (5), the causal effect of X on Y is given by

$$pr(y|\check{x}) = \sum_u pr(y|x, u)pr(u) \quad . \tag{8}$$

Using the conditional independence assumptions implied by the decomposition (7), we can eliminate u from (8) to obtain:

$$pr(y|\check{x}) = \sum_z pr(z|x) \sum_{x'} pr(y|x', z)pr(x') \tag{9}$$

We summarize this result by a theorem.

THEOREM 2. If a set of variables Z satisfies the following conditions relative to an ordered pair of variables (X, Y): (i) Z intercepts all direct paths from X to Y, (ii) There is no back-door path between X and Z, (iii) Every back-door path between Z and Y is blocked by X, then the causal effect of X on Y is identifiable and is given by (9).

The graphical criterion of Theorem 2 uncovers many new structures that permit the identification of causal effects from measurements of variables that are affected by treatments: see Section 5.1. 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 some, also induces an inborn craving for nicotine.

In this case, (9) would provide us with the means to quantify, from nonexperimental data, the causal effect of smoking on cancer, assuming, of course, that $pr(x, y, z)$ is made available and that we believe that smoking does not have any direct effect on lung cancer except that mediated by tar deposits.

4 A CALCULUS OF INTERVENTION

This section establishes a set of inference rules by which probabilistic sentences involving interventions 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 $pr(y|\check{x})$, where X and Y stand for any subsets of observed variables. By derivation we mean step-wise reduction of the expression $pr(y|\check{x})$ 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 4.

4.1 *Preliminary notation*

Let X, Y, and Z be arbitrary disjoint sets of nodes in a directed acyclic graph 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, $pr(y|\check{x}, z) \triangleq pr(y, z|\check{x})/pr(z|\check{x})$ denotes the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x.

4.2 *Inference rules*

The following theorem states the three basic inference rules of the proposed calculus. Proofs are provided in the appendix.

THEOREM 3. Let G be the directed acyclic graph associated with a causal model as defined in (3), and let $pr(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z, and W we have:

Rule 1 Insertion/deletion of observations:

$$pr(y|\check{x}, z, w) = pr(y|\check{x}, w) \;\text{ if }\; (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}. \tag{10}$$

Rule 2 Action/observation exchange:

$$pr(y|\check{x}, \check{z}, w) = pr(y|\check{x}, z, w) \;\text{ if }\; (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}. \tag{11}$$

Rule 3 Insertion/deletion of actions:

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

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 "$\check{x}$" operator as a replacement of the causal mechanism that connects X to its pre-intervention 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 (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.

COROLLARY 1. A causal effect $q = pr(y_1, ..., y_k|\check{x}_1, ..., \check{x}_m)$ 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 3, which reduces q into a standard i.e., hat-free, probability expression involving observed quantities. $\square$

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 as described in an unpublished report (R-226, CSD, University of California, Los Angeles, 1994) by D. Galles. As the next subsection illustrates, symbolic derivations using the hat notation are much more convenient than algebraic derivations that aim at eliminating 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 Rules 1-3 can be used to derive all causal effect estimands in the structure of Figure 3 above. Figure 4 displays the subgraphs that will be needed for the derivations that follow.

Task 1, compute $pr(z|\check{x})$

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

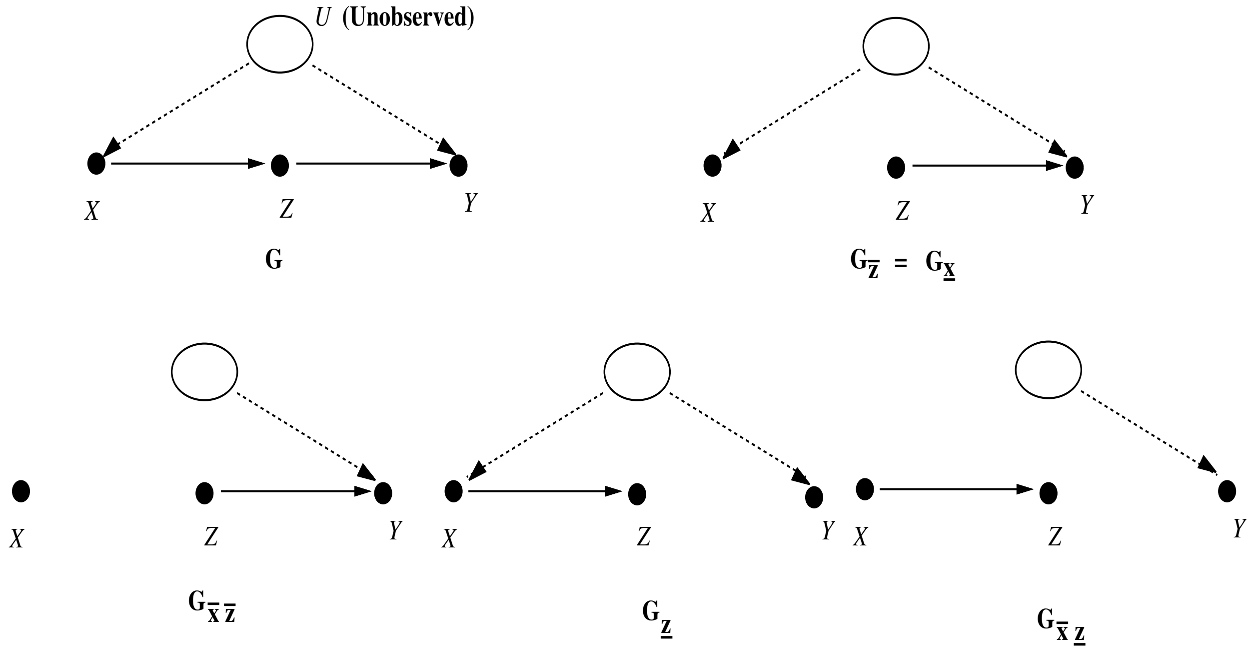

Figure 4: Subgraphs of G used in the derivation of causal effects.

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

$$pr(z|\check{x}) = pr(z|x). \tag{13}$$

Task 2, compute $pr(y|\check{z})$.
Here we cannot apply Rule 2 to exchange $\check{z}$ 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 measuring variables, such as X, that reside on that path. This involves conditioning and summing over all values of X:

$$pr(y|\check{z}) = \sum_x pr(y|x, \check{z}) pr(x|\check{z}) \tag{14}$$

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

$$pr(x|\check{z}) = pr(x) \quad \text{if} \quad (Z \perp\!\!\!\perp X)_{G_{\overline{Z}}} \tag{15}$$

since X and Z are d-separated in $G_{\overline{Z}}$. Intuitively, manipulating Z should have no effect on X, because Z is a descendant of X in G. To reduce the former, $pr(y|x, \check{z})$, we consult Rule 2:

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

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

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

which is a special case of the back-door formula (6). The legitimizing condition, $(Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}}$, offers yet another graphical test for the ignorability condition of Rosenbaum & Rubin (1983).

Task 3, compute $pr(y|\check{x})$.

Writing

$$pr(y|\check{x}) = \sum_z pr(y|z, \check{x}) pr(z|\check{x}) \tag{18}$$

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

$$pr(y|z, \check{x}) = pr(y|\check{z}, \check{x}) \tag{19}$$

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

$$pr(y|z, \check{x}) = pr(y|\check{z}) \tag{20}$$

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

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

which is identical to the front-door formula (9).

The reader can easily verify that all other causal effects, for example, $pr(y, z|\check{x})$ and $pr(x, z|\check{y})$, can likewise be derived through the rules of Theorem 3. Note that in all the derivations the graph G provides 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 $pr(y|\check{x})$ is not identifiable and, for practical reasons of cost or ethics, we cannot control X by randomized experiment. The question arises whether $pr(y|\check{x})$ 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 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 $pr(y|\check{x})$ into expressions in which only members of Z obtain the hat symbol. Using Theorem 3 it can be shown that the following conditions are sufficient for admitting a surrogate variable Z: (i) X intercepts all directed paths from Z to Y, and, (ii) $pr(y|\check{x})$ is identifiable in $G_{\overline{Z}}$. Indeed, if condition (i) holds, we can write $pr(y|\check{x}) = pr(y|\check{x}, \check{z})$, because $(Y \perp\!\!\!\perp Z|X)_{G_{\overline{XZ}}}$. But $pr(y|\check{x}, \check{z})$ stands for the causal effect of X on Y in a model governed by $G_{\overline{Z}}$ which, by condition (ii), is identifiable. Figures 7(e) and 7(h) below illustrate models in which both conditions hold. Translated to our cholesterol example, these condition require 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

Figure 5 shows simple diagrams in which $pr(y|\check{x})$ cannot be identified due to the presence of a bow pattern, i.e., a confounding arc (dashed) embracing a causal link between X and Y. 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 $pr(y|\check{x})$ 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 $pr(y|\check{x})$ identifiable. For example, if Y is related to X via a linear relation $Y = bX + U$, where U is an unobserved disturbance possibly correlated with X, then $b = \frac{\partial}{\partial x} E(Y|\check{x})$ 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 $E(Y|\check{x})$ via the instrumental-variable formula (Bowden & Turkington, 1984; Angrist et al., 1993):

$$b \triangleq \frac{\partial}{\partial x} E(Y|\check{x}) = \frac{E(Y|z)}{E(X|z)} = \frac{R_{yz}}{R_{xz}}. \tag{22}$$

In nonparametric models, adding an instrumental variable Z to a bow-pattern (Figure 5(b)) does not permit the identification of $pr(y|\check{x})$. 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 $pr(y|\check{x})$ without making additional assumptions on the nature of the interactions between compliance and response (Imbens & Angrist, 1994), as is done, for example, in the general approach to instrumental variables developed in Angrist et al. (1993). While the added arc $Z \rightarrow X$ permits us to calculate bounds on $pr(y|\check{x})$ (Robins, 1989, Sec. 1g; Manski, 1990) and the upper and lower bounds may even coincide for certain types of distributions $pr(x, y, z)$ (Balke & Pearl, 1994), there is no way of computing $pr(y|\check{x})$ for every positive distribution $pr(x, y, z)$, as required by Definition 4.

In general, the addition of arcs to a causal diagram can impede, but never assist, the identification of causal effects in nonparametric models. 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 1, would succeed in the original diagram.

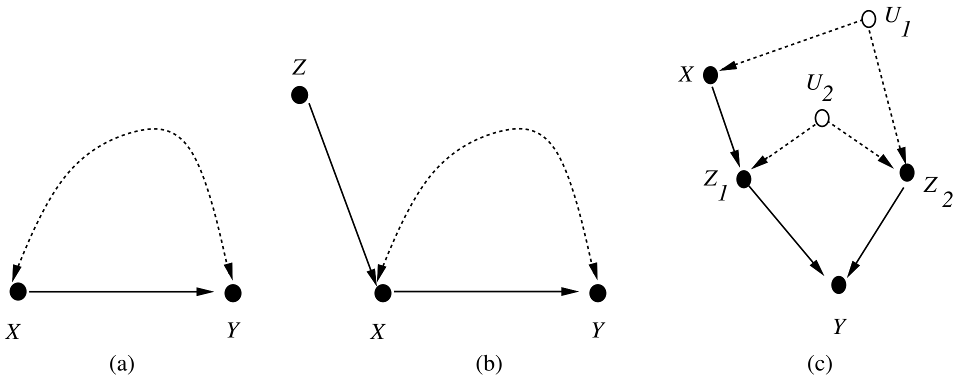

Figure 5: (a) A bow-pattern: a confounding arc embracing a causal link $X \rightarrow Y$, thus preventing the identification of $pr(y|\check{x})$ even in the presence of an instrumental variable Z, as in (b). (c) A bow-less graph still prohibiting the identification of $pr(y|\check{x})$.

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

5.1 *Identifying models*

Figure 6 shows simple diagrams in which the causal effect of X on Y is identifiable. Such models are called identifying because their structures communicate a sufficient number of assumptions to permit the identification of the target quantity $pr(y|\check{x})$. Latent variables are not shown explicitly in these diagrams; rather, such variables are implicit in the confounding arcs (dashed). Every causal diagram with latent variables can be converted to an equivalent diagram involving measured variables interconnected by arrows and confounding arcs. This conversion corresponds to substituting out all latent variables from the structural equations of (3) and then constructing a new diagram by connecting any two variables X_i and X_j by (i) an arrow from X_j to X_i whenever X_j appears in the equation for X_i and (ii) a confounding arc whenever the same ϵ term appears in both f_i and f_j. The result is a diagram in which all unmeasured variables are exogenous and mutually independent.

Several features should be noted from examining the diagrams in Figure 6.

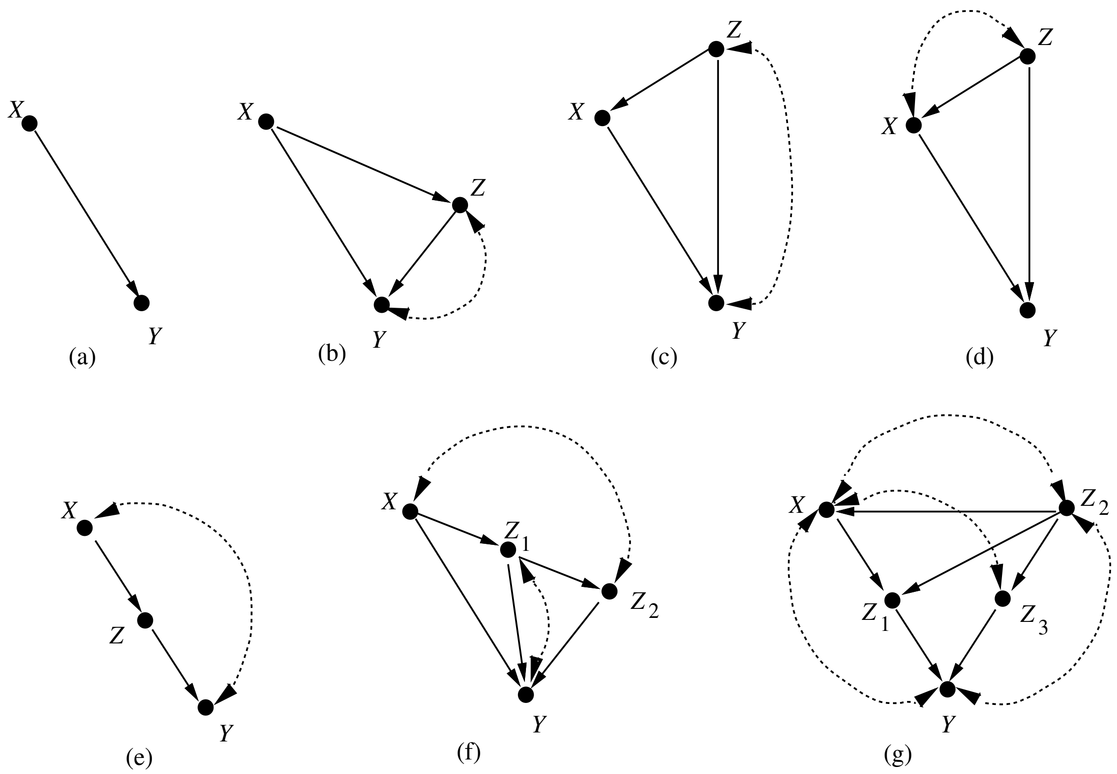

Figure 6: Typical models in which the effect of X on Y is identifiable. Dashed arcs represent confounding paths, and Z represents observed covariates.

(i) Since the removal of any arc or arrow from a causal diagram can only assist the identifiability of causal effects, $pr(y|\check{x})$ will still be identified in any edge-subgraph of the diagrams shown in Figure 6. Likewise, the introduction of mediating observed variables onto any edge in a causal graph can assist, but never impede, the identifiability of any causal effect. Therefore, $pr(y|\check{x})$ will still be identified from any graph obtained by adding mediating nodes to the diagrams shown in Figure 6.

(ii) The diagrams in Figure 6 are maximal, in the sense that the introduction of any additional arc or arrow onto an existing pair of nodes would render $pr(y|\check{x})$ no longer identifiable.

(iii) Although most of the diagrams in Figure 6 contain bow-patterns, none of these patterns emanates from X as is the case in Figure 7(a) and (b) below. In general, a necessary condition for the identifiability of $pr(y|\check{x})$ is the absence of a confounding arc between X and any child of X that is an ancestor of Y.

(iv) Diagrams (a) and (b) in Figure 6 contain no back-door paths between X and Y, and thus represent experimental designs in which there is no confounding bias between the treatment (X) and the response (Y), i.e., X is strongly ignorable relative to Y (Rosenbaum & Rubin, 1983); hence, $pr(y|\check{x}) = pr(y|x)$. Likewise, diagrams (c) and (d) in Figure 6 represent designs in which observed covariates, Z, block every back-door path between X and Y, i.e., X is conditionally ignorable given Z (Rosenbaum & Rubin, 1983); hence, $pr(y|\check{x})$ is obtained by standard adjustment for Z, as in (6):

$$pr(y|\check{x}) = \sum_z pr(y|x,z)pr(z).$$

(v) For each of the diagrams in Figure 6, we can readily obtain a formula for $pr(y|\check{x})$, by using symbolic derivations patterned after those in Section 4.3. The derivation is often guided by the graph topology. For example, diagram (f) in Figure 6 dictates the following derivation. Writing

$$pr(y|\check{x}) = \sum_{z_1,z_2} pr(y|z_1,z_2,\check{x})pr(z_1,z_2|\check{x})$$

we see that the subgraph containing $\{X, Z_1, Z_2\}$ is identical in structure to that of diagram (e), with (Z_1, Z_2) replacing (Z, Y), respectively. Thus, $pr(z_1, z_2|\check{x})$ can be obtained from (14) and (21). Likewise, the term $pr(y|z_1, z_2, \check{x})$ can be reduced to $pr(y|z_1, z_2, x)$ by Rule 2, since $(Y \perp\!\!\!\perp X|Z_1, Z_2)_{G_{\underline{X}}}$. Thus, we have

$$pr(y|\check{x}) = \sum_{z_1,z_2} pr(y|z_1,z_2,x)\ pr(z_1|x) \sum_{x'} pr(z_2|z_1,x')\ pr(x'). \tag{23}$$

Applying a similar derivation to diagram (g) of Figure 6 yields

$$pr(y|\check{x}) = \sum_{z_1}\sum_{z_2}\sum_{x'} pr(y|z_1,\ z_2,\ x')pr(x')pr(z_1|z_2,\ x)pr(z_2). \tag{24}$$

Note that the variable Z_3 does not appear in the expression above, which means that Z_3 need not be measured if all one wants to learn is the causal effect of X on Y.

(vi) In diagrams (e), (f), and (g) of Figure 6, the identifiability of $pr(y|\check{x})$ is rendered feasible through observed covariates, Z, that are affected by the treatment X, i.e., descendants of X. This stands contrary to the warning, repeated in most of the literature on statistical experimentation, to refrain from adjusting for concomitant observations that are affected by the treatment (Cox, 1958; Rosenbaum, 1984; Pratt & Schlaifer, 1988; Wainer, 1989). It is commonly believed that if a concomitant Z is affected by the treatment, then it must be excluded from the analysis of the total effect of the treatment (Pratt & Schlaifer, 1988). The reason given for the exclusion is that the calculation of total effects amounts to integrating out Z, which is functionally equivalent to omitting Z to begin with. Diagrams (e), (f), and (g) show cases where one wants to learn the total effects of X and, still, the measurement of concomitants that are affected by X (e.g., Z, or Z_1) is necessary. However, the adjustment needed for such concomitants is nonstandard, involving two or more stages of the standard adjustment of (6): see (9), (23), and (24)).

(vii) In diagrams (b), (c), and (f) of Figure 6, Y has a parent whose effect on Y is not identifiable yet the effect of X on Y is identifiable. This demonstrates that local identifiability is not a necessary condition for global identifiability. In other words, to identify the effect of X on Y we need not insist on identifying each and every link along the paths from X to Y.

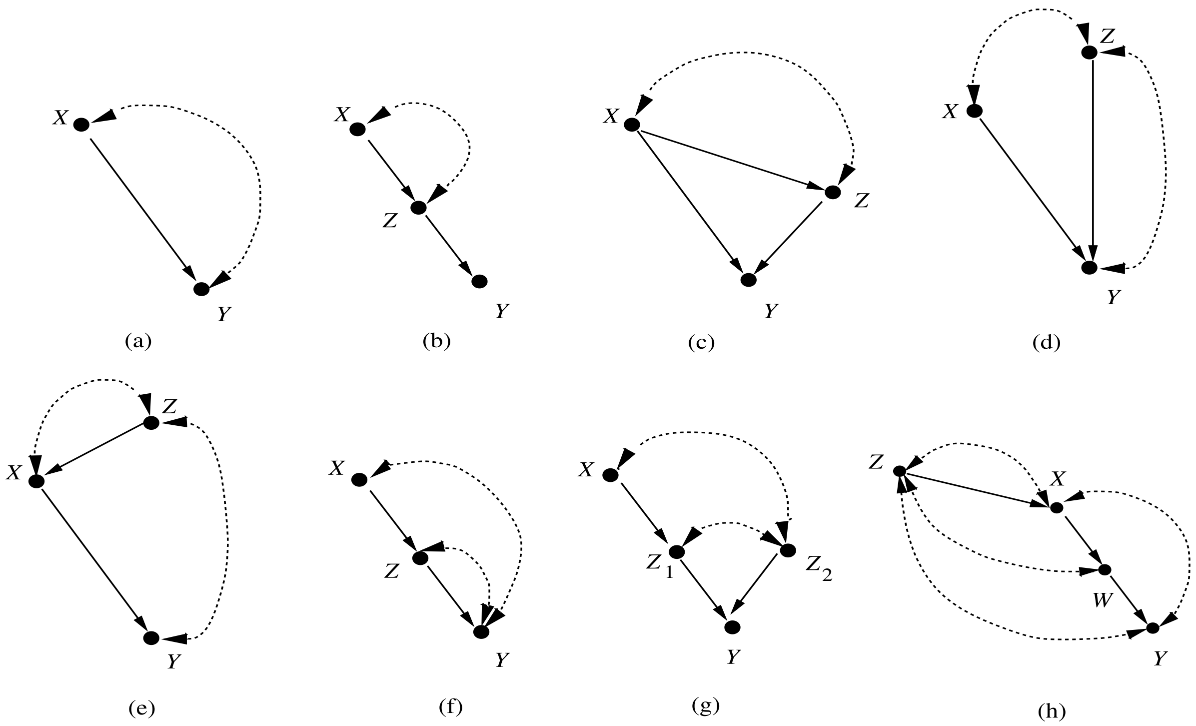

Figure 7: Typical models in which $pr(y|\check{x})$ is not identifiable.

5.2 *Nonidentifying models*

Figure 7 presents typical diagrams in which the total effect of X on Y, $pr(y|\check{x})$, is not identifiable. Noteworthy features of these diagrams are as follows.

(i) All graphs in Figure 7 contain unblockable back-door paths between X and Y, that is, paths ending with arrows pointing to X which cannot be blocked by observed nondescendants of X. The presence of such a path in a graph is, indeed, a necessary test for nonidentifiability. It is not a sufficient test, though, as is demonstrated by Figure 6(e), in which the back-door path (dashed) is unblockable and yet $pr(y|\check{x})$ is identifiable.

(ii) A sufficient condition for the nonidentifiability of $pr(y|\check{x})$ is the existence of a confounding path between X and any of its children on a path from X to Y, as shown in Figure 7(b) and (c). A stronger sufficient condition is that the graph contain any of the patterns shown in Figure 7 as an edge-subgraph.

(iii) Graph (g) in Figure 7 demonstrates that local identifiability is not sufficient for global identifiability. For example, we can identify $pr(z_1|\check{x})$, $pr(z_2|\check{x})$, $pr(y,|\check{z}_1)$, and $pr(y|\check{z}_2)$, but not $pr(y|\check{x})$. This is one of the main differences between nonparametric and linear models; in the latter, all causal effects can be determined from the structural coefficients, each coefficient representing the causal effect of one variable on its immediate successor.

6 DISCUSSION

The basic limitation of the methods proposed in this paper is that the results must rest on the causal assumptions shown in the graph, and that these cannot usually be tested in observational studies. In related papers (Pearl, 1994a, 1995) we show that some of the assumptions, most notably those associated with instrumental variables (Figure 5 (b)) are subject to falsification tests. Additionally, considering that any causal inferences from observational studies must ultimately rely on some kind of causal assumptions, the methods described in this paper offer an effective language for making those assumptions precise and explicit, so they can be isolated for deliberation or experimentation and, once validated, be integrated with statistical data.

A second limitation concerns an assumption inherent in identification analysis, namely, that the sample size is so large that sampling variability may be ignored. The mathematical derivation of causal-effect estimands should therefore be considered a first step toward supplementing estimates of these with confidence intervals and significance levels, as in traditional analysis of controlled experiments. Having nonparametric estimates for causal effects does not imply that one should refrain from using parametric forms in the estimation phase of the study. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand in (9) can be replaced by $E(Y|\check{x}) = R_{xz}\beta_{zy \cdot x}x$ where $\beta_{zy \cdot x}$ is the standardized regression coefficient, and the estimation problem reduces to that of estimating regression coefficients. More sophisticated estimation techniques, can be found in Rubin (1978), Robins (1989, Sec. 17), and Robins et al. (1992, pp. 331–333).

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 set X of treatment variables is made to respond in a specified way to some set Z of covariates, 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 $pr(y|\check{x}, z)$.

A second extension concerns the use of the intervention calculus of Theorem 3 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) 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). Secondly, the d-separation criterion for directed acyclic graphs 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 $pr(y|\check{x})$ to hat-free expressions may require the solution of nonlinear equations.

Finally, a few comments regarding the notation introduced in this paper. There have been three approaches to expressing causal assumptions in mathematical form. The most common approach in the statistical literature invokes the Rubin's model (Rubin, 1974) in which probability functions are defined over an augmented space of observable and counterfactual variables. In this model, causal assumptions are expressed as independence constraints over the augmented probability function, as exemplified by Rosenbaum and Rubin's (1983) definitions of ignorability conditions. An alternative but related approach, still using the standard language of probability, is to define augmented probability functions over variables representing hypothetical interventions (Pearl, 1993b).

The language of structural models, which includes path diagrams (Wright, 1921) and structural equations (Goldberger, 1974) represents a drastic departure from these two approaches, because it invokes new primitives, such as arrows, disturbance terms, or plain causal statements, which have no parallels in the language of probability. This language has been very popular in the social sciences and econometrics, because it closely echoes statements made in ordinary scientific discourse and thus provides a natural way for scientists to communicate knowledge and experience, especially in situations involving many variables.

Statisticians, however, have generally found structural models suspect, because the empirical content of basic notions in these models appears to escape conventional methods of explication. For example, analysts have found it hard to conceive of experiments, however hypothetical, whose outcomes would be constrained by a given structural equation. Standard probability calculus 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 this equation certain variables that are highly correlated with X or Y. As a consequence, the whole enterprise of structural equation modeling has become the object of serious controversy and misunderstanding among researchers (Freedman, 1987; Wermuth, 1992; Whittaker, 1990, p. 302; Cox & Wermuth, 1993).

To a large extent, this history of controversy stems not from faults in the structural modeling approach but, rather, from a basic limitation of standard probability theory: when viewed as a mathematical language, it is too weak to describe the precise experimental conditions that prevail in a given study. For example, standard probabilistic notation cannot distinguish between an experiment in which variable X is observed to take on value x and one in which variable X is set to value x by some external control. The need for this distinction was recognized by several researchers, most notably Pratt & Schlaifer (1988) and Cox (1992), but has not led to a more refined and manageable mathematical notation capable of reflecting this distinction.

The "hat" notation developed in this paper permits one to specify precisely what is being held constant and what is merely measured in a given study and, using this specification, the basic notions of structural models can be given clear empirical interpretation. For example, the meaning of b in the equation $Y = bX + \epsilon_Y$ is simply $\frac{\partial}{\partial x} E(Y|\check{x})$, 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, for example, via another equation: $X = aY + \epsilon_X$. Moreover, the notion of randomization need not be invoked. Likewise, the analyst's decision as to which variables should be included in a given equation is based on a hypothetical controlled experiment: A variable Z is excluded from the equation for Y if it has no influence on Y when all other variables, S_{YZ}, are held constant, that is, $pr(y|\check{z}, \check{s}_{YZ}) = pr(y|\check{s}_{YZ})$. Specifically, variables that are excluded from the equation $Y = bX + \epsilon_Y$ are not conditionally independent of Y given measurements of X, but rather causally irrelevant to Y given settings of X. The operational meaning of the so-called "disturbance term", ϵ_Y, is likewise demystified: ϵ_Y is defined as the difference $Y - E(Y|\check{s}_Y)$; two disturbance terms, ϵ_X and ϵ_Y, are correlated if $pr(y|\check{x}, \check{s}_{XY}) \neq pr(y|x, \check{s}_{XY})$; and so on.

The distinctions provided by the "hat" notation clarify the empirical basis of structural equations and should make causal models more acceptable to empirical researchers. 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 an effective means for scientists to communicate subject-matter information, and to infer its logical consequences when combined with statistical data.

APPENDIX (Proof of Theorem 3)

(i) Rule 1 follows from the fact that deleting equations from the model in Eq. (8) results, again, in a recursive set of equations in which all ϵ terms are mutually independent. The d-separation condition is valid for any recursive model, hence it is valid for the submodel resulting from deleting the equations for X. Finally, since the graph characterizing this submodel is given by $G_{\overline{X}}$, $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$ implies the conditional independence $pr(y|\check{x}, z, w) = pr(y|\check{x}, w)$ in the post-intervention distribution.

(ii) The graph $G_{\overline{X}\underline{Z}}$ differs from $G_{\overline{X}}$ only in lacking the arrows emanating from Z, hence it retains all the back-door paths from Z to Y that can be found in $G_{\overline{X}}$. The condition $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$ ensures that all back-door paths from Z to Y (in $G_{\overline{X}}$) are blocked by $\{X, W\}$. Under such conditions, setting $Z = z$ or conditioning on $Z = z$ has the same effect on Y. This can best be seen from the augmented diagram $G'_{\overline{X}}$, to which the intervention arcs $F_Z \to Z$ were added, where F_z stands of the functions that determine Z in the structural equations (Pearl, 1993b). If all back-door paths from F_Z to Y are blocked, the remaining paths from F_Z to Y must go through the children of Z, hence these paths will be blocked by Z. The implication is that Y is independent of F_Z given Z, which means that the observation $Z = z$ cannot be distinguished from the intervention $F_Z = set(z)$.

(iii) The following argument was developed by D. Galles. Consider the augmented diagram $G'_{\overline{X}}$ to which the intervention arcs $F_z \rightarrow Z$ are added. If $(F_Z \perp\!\!\!\perp Y|W,X)_{G'_{\overline{X}}}$, then $pr(y|\check{x},\check{z},w) = pr(y|\check{x},w)$. If $(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\ \overline{Z(W)}}}$, and $(F_Z \not\perp\!\!\!\perp Y|W,X)_{G'_{\overline{X}}}$, there must be an unblocked path from a member $F_{Z'}$ of F_Z to Y that passes either through a head-to-tail junction at Z', or a head-to-head junction at Z'. If there is such a path, let P be the shortest such path. We will show that P will violate some premise, or there exists a shorter path, either of which leads to a contradiction.

If the junction is head-to-tail, that means that $(Y \not\perp\!\!\!\perp Z'|W,X)_{G'_{\overline{X}}}$, but $(Y \perp\!\!\!\perp Z'|W,X)_{G'_{\overline{X}\ \overline{Z(W)}}}$. So, there must be an unblocked path from Y to Z' that passes through some member Z'' of $Z(W)$ in either a head-to-head or a tail-to-head junction. This is impossible. If the junction is head-to-head, then some descendant of Z'' must be in W for the path to be unblocked, but then Z'' would not be in $Z(W)$. If the junction is tail-to-head, there are two options : either the path from Z' to Z'' ends in a arrow pointing to Z'', or an arrow pointing away from Z''. If it ends in an arrow pointing away from Z'', then there must be a head-to-head junction along the path from Z' to Z''. In that case, for the path to be unblocked, W must be a descendant of Z'', but then Z'' would not be in $Z(W)$. If it ends in an arrow pointing to Z'', then there must be an unblocked path from Z'' to Y in $G_{\overline{X}}$ that is blocked in $G_{\overline{X}\ \overline{Z(W)}}$. If this is true, then there is an unblocked path from $F_{Z''}$ to Y that is shorter than P, the shortest path.

If the junction through Z' is head-to-head, then either Z' is in $Z(W)$, in which case that junction would be blocked, or there is an unblocked path from Z' to Y in $G_{\overline{X},\ \overline{Z(W)}}$ that is blocked in $G_{\overline{X}}$. Above, we proved that this could not occur.

So $(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\ \overline{Z(W)}}}$ implies $(F_Z \perp\!\!\!\perp Y|W,X)_{G'_{\overline{X}}}$, and thus $pr(y|\check{x},\check{z},w) = pr(y|\check{x},w)$.

APPENDIX II: Graphs, structural equations, and counterfactuals

This paper uses two representations of causal models: graphs and structural equations. By now, both representations have been considered controversial for almost a century. On the one hand, economists and social scientists have embraced these modeling tools, but they continue to debate the empirical content of the symbols they estimate and manipulate; as a result, the use of structural models in policy-making contexts is often viewed with suspicion. Statisticians, on the other hand, reject both representations as problematic (if not meaningless) and instead resort to the Neyman-Rubin counterfactual notation whenever they are pressed to communicate causal information. This appendix presents an explication that unifies these three representation schemes in order to uncover commonalities, mediate differences, and make the causal-inference literature more generally accessible.

The primitive object of analysis in Rubin's counterfactual framework is the unit-based response variable, denoted $Y(x,u)$ or $Y_x(u)$, read: "the value that Y would obtain in unit u, had X been x". This variable has natural interpretation in structural equations models. Consider a set T of equations

$$X_i = f_i(PA_i, U_i) \quad i = 1, \ldots, n \tag{25}$$

where the U_i stand for latent exogenous variables (or disturbances), and the PA_i are the explanatory (observed) variables in the ith equation (pa_i is a realization of PA_i). (25) is similar to (3), except we no longer insist on the equations being recursive or on the U_i's being independent. Let U stand for the vector $(U_1, \ldots, U_n)$, let X and Y be two disjoint subsets of observed variables, and let T_x be the submodel created by replacing the equations corresponding to variables in X with $X = x$, as in Definition 2. The structural interpretation of $Y(x, u)$ is given by

$$Y(x, u) \triangleq Y_{T_x}(u) \tag{26}$$

namely, $Y(x, u)$ is the (unique) solution of Y under the realization $U = u$ in the submodel T_x of T. While the term *unit* in the counterfactual literature normally stands for the identity of a specific individual in a population, a unit may also be thought of as the set of attributes that characterize that individual, the experimental conditions under study, the time of day, and so on, which are represented as components of the vector u in structural modeling. Eq. (26) forms a connection between the opaque English phrase "the value that Y would obtain in unit u, had X been x" and the physical processes that transfer changes in X into changes in Y. The formation of the submodel T_x represents a minimal change in model T needed for making x and u compatible; such a change could result either from external intervention or from a natural yet unanticipated eventuality.

Given this interpretation of $Y(x, u)$, it is instructive to contrast the methodologies of causal inference in the counterfactual and the structural frameworks. If U is treated as a random variable, then the value of the counterfactual $Y(x, u)$ becomes a random variable as well, denoted as $Y(x)$ or Y_x. The counterfactual analysis proceeds by imagining the observed distribution $P^*(x_1, \ldots, x_n)$ as the marginal distribution of an augmented probability function P^* defined over both observed and counterfactual variables. Queries about causal effects, written $P(y|\hat{x})$ in the structural analysis, are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $P^*(Y(x) = y)$. The new entities $Y(x)$ are treated as ordinary random variables that are connected to the observed variables via consistency constraints (Robins, 1987) such as

$$X = x \implies Y(x) = Y \tag{27}$$

and a set of conditional independence assumptions which the investigator must supply to endow the augmented probability, P^*, with causal knowledge, paralleling the knowledge that a structural analyst would encode in equations or in graphs.

For example, to communicate the understanding that in a randomized clinical trial (see Figure 6(b)) the way subjects react (Y) to treatments (X) is statistically independent of the treatment assignment (Z), the analyst would write $Y(x) \perp\!\!\!\perp Z$. Likewise, to convey the understanding that the assignment processes is randomized, hence independent of any variation in the treatment selection process, structurally written $U_X \perp\!\!\!\perp U_Z$, the analyst would use the independence constraint $X(z) \perp\!\!\!\perp Z$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest, for example, $P^*(Y(x) = y)$; in other cases, only bounds on the solution can be obtained. Section 6 explains why this approach is conceptually appealing to some statisticians, even though the process of eliciting judgments about counterfactual dependencies has so far not been systematized. When counterfactual variables are not viewed as by-products of a deeper, process-based model, it is hard to ascertain whether *all* relevant judgments have been articulated, whether the judgments articulated are redundant, or whether those judgments are self-consistent. The elicitation of such judgments can be systematized using the following translation from graphs.

Graphs provide qualitative information about the structure of both the equations in the model and the probability function $P(u)$, the former is encoded as missing arrows, the latter as missing dashed arcs. Each parent-child family (PA_i, X_i) in a causal diagram G corresponds to an equation in the model (25). Hence, missing arrows encode exclusion assumptions, that is, claims that adding excluded variables to an equation will not change the outcome of the hypothetical experiment described by that equation. Missing dashed arcs encode independencies among disturbance terms in two or more equations. For example, the absence of dashed arcs between a node Y and a set of nodes $Z_1, \ldots, Z_k$ implies that the corresponding error variables, $U_Y, U_{Z_1}, \ldots, U_{Z_k}$, are jointly independent in $P(u)$.

These assumptions can be translated into the counterfactual notation using two simple rules; the first interprets the missing arrows in the graph, the second, the missing dashed arcs.

1. Exclusion restrictions: For every variable Y having parents PA_Y, and for every set of variables S disjoint of PA_Y, we have
$$Y(pa_Y) = Y(pa_Y, s) \tag{28}$$
2. Independence restrictions: If $Z_1, \ldots, Z_k$ is any set of nodes not connected to Y via dashed arcs, we have
$$Y(pa_Y) \perp\!\!\!\perp \{Z_1(pa_{Z_1}), \ldots, Z_k(pa_{Z_k})\} \tag{29}$$

Given a sufficient number of such restrictions on P^*, it is possible to compute causal effects $P^*(Y(x) = y)$ using standard probability calculus together with the logical constraints (e.g., Eq. (27) that couple counterfactual variables with their measurable counterparts. These constraints can be used as axioms, or rules of inference, in attempting to transform causal effect expressions, $P^*(Y(x) = y)$, into expressions involving only measurable variables. When such a transformation is found, the corresponding causal effect is identifiable, since P^* reduces then to P. The axioms needed for such transformation are:

$$\begin{aligned}
&\textbf{Degeneracy}: \quad Y(\emptyset) = Y && (30)\\
&\textbf{Composition}: \quad Y(x) = Y(x, Z(x)) \text{ for any } Z \text{ disjoint of } \{X, Y\} && (31)\\
&\textbf{Sure-thing}: \quad \text{If } Y(x,z) = Y(x',z)\ \forall\, x' \neq x, \text{ then } Y(x,z) = Y(z) && (32)
\end{aligned}$$

Degeneracy asserts that the observed value of Y is equivalent to a counterfactual variable $Y(x)$ in which the conditional part: "had X been x" is not enforced, that is, X is the empty set.

The Composition axiom asserts:

$$\text{If } Y(x,z) = y \text{ and } Z(x) = z, \text{ then } Y(x) = y$$

and, conversely:

$$\text{If } Y(x) = y \text{ and } Z(x) = z, \text{ then } Y(x,z) = y$$

In words: "The value that Y would obtain had X been x is the same as that obtained had X been x and Z been z, where z is the value that Z would obtain had X been x".

The sure-thing axiom (named after Savage's "sure-thing principle") asserts that if $Y(x,z) = y$ for every value x of X, then the counterfactual antecedant $X = x$ is redundant, namely, we need not concern ourselves with the value that X actually obtains.

Properties (31)-(32) are theorems in the structural interpretation of $Y(x,u)$ as given in Eq. (26). However, in the Neyman-Rubin model, where $Y(x,u)$ is taken as a primitive notion, these properties must be considered axioms which, together with other such properties, defines the abstract counterfactual conditioning operator "had X been x".[1] It is easy to verify that composition and degeneracy imply the consistency rule of (27); substituting $X = \{\emptyset\}$ in (34) yields $Y = Y(z)$ if $Z = z$, which is equivalent to (27).

As an example, let us compute the causal effects associated with the model shown in Figure 4. The parents sets a given by:

$$PA_X = \{\emptyset\},\ PA_Z = \{X\},\ PA_Y = \{Z\} \tag{33}$$

Consequently, the exclusion restrictions translate into:

$$\begin{aligned}
Z(x) &= Z(y,x) && (34)\\
X(y) &= X(z,y) = X(z) = X && (35)\\
Y(z) &= Y(z,x) && (36)
\end{aligned}$$

The absence of a dashed arc between Z and $\{Y, X\}$ translates into the independence restrictions:

$$Z(x) \perp\!\!\!\perp \{Y(z), X\} \tag{37}$$

[1] The composition rule was communicated to me by James Robins (1995, in conversation) as a property needed for defining graph models (called "finest fully randomized causal graphs" in (Robins pp. 1419-1423, 1986)); in Robins' treatment of counterfactuals, $Y(x,z)$ and $Z(x)$ may not be defined.

Task-1, compute $P^*(Z(x)=z)$ (Equivalently $P(z|\hat{x})$)
From (37) we have $Z(x) \perp\!\!\!\perp X$, hence

$$P^*(Z(x)=z) = P^*(Z(x)=z|x) = P^*(z|x) = P(z|x) \tag{38}$$

Task-2, compute $P^*(Y(z)=y)$ (Equivalently $P^*(y|\hat{z})$)

$$P^*(Y(z)=y) = \sum_x P^*(Y(z)=y|x)P^*(x) \tag{39}$$

From (37) we have

$$Y(z) \perp\!\!\!\perp Z(x)|X \tag{40}$$

hence

$$\begin{aligned} P^*(Y(z)=y|x) &= P^*(Y(z)=y|x, Z(x)=z) && \text{by (52)} \\ &= P^*(Y(z)=y|x,z) && \text{by (40)} \\ &= P^*(y|x,z) && \text{by (40)} \\ &= P(y|x,z) \end{aligned} \tag{41}$$

Substituting (41) in (39), gives

$$P^*(Y(z)=y) = \sum_x P(y|x,z)P(x) \tag{42}$$

which is the celebrated covariate-adjustment formula for causal effect, as in Eq. (6).

Task-3, compute $P^*(Y(x)=y)$ (Equivalently $P(y|\hat{x})$)
For any arbitrary variable Z, we have (by composition)

$$Y(x) = Y(x, Z(x))$$

In particular, since $Y(x,z) = Y(z)$ (from (36)), we have

$$Y(x) = Y(x, Z(x)) = Y(Z(x))$$

and

$$\begin{aligned} P^*(Y(x)=y) &= P^*(Y(Z(x))=y) \\ &= \sum_z P^*(Y(Z(x))=y)|Z(x)=z)\ P^*(Z(x)=z) \\ &= \sum_z P^*(Y(z)=y)|Z(x)=z)\ P^*(Z(x)=z) \\ &= \sum_z P^*(Y(z)=y)\ P^*(Z(x)=z) \end{aligned}$$

since $Y(z) \perp\!\!\!\perp Z(x)$.

$P^*(Y(z)=y)$ and $P^*(Z(x)=z)$ were computed in (42) and (38), respectively, hence

$$P^*(Y(x)=y) = \sum_z P(z|x) \sum_{x'} P(y|z,x')P(z')$$

In summary, the structural and counterfactual frameworks are complementary of each other. Structural analysts can interpret counterfactual sentences as constraints over the solution set of a given system of equations (26) and, conversely, counterfactual analysts can use the constraints (over P^*) given by Eqs. (28) and (29) as a definition of graphs, structural equations and the physical processes which they represent.

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, Arthur Dempster, David Galles, Arthur Goldberger, Sander Greenland, David Hendry, Paul Holland, Guido Imbens, Ed Leamer, Rod McDonald, John Pratt, Paul Rosenbaum, Keunkwan Ryu, Glenn Shafer, Michael Sobel, David Tritchler, and Nanny Wermuth. The research was partially supported by grants from AFOSR and NSF.

REFERENCES

Angrist, J.D., Imbens, G.W., & Rubin, D.B. (1993). Identification of causal effects using instrumental variables. To appear in *JASA*.

Balke, A., & Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence*, Eds R. Lopez de Mantaras and D. Poole, 46–54. Morgan Kaufmann: San Mateo, CA.

Bowden, R.J., & Turkington, D.A. (1984). *Instrumental Variables*, Cambridge University Press: Cambridge, MA.

Cox, D.R. (1958). *The Planning of Experiments*, John Wiley and Sons: New York.

Cox, D.R. (1992). Some statistical aspects. *Journal of the Royal Statistical Society, Series A* **155**, 291–301.

Cox, D.R., & Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science* **8**, 204–218.

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

Fisher, F.M. (1970). A correspondence principle for simultaneous equation models. *Econometrica* **38**, 73–92.

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

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

Geiger, D., Verma, T.S., & Pearl, J. (1990). Identifying in dependence in Bayesian networks. *Networks* **20**, 507–534.

Goldberger, A.S. (1973). *Structural Equation Models in the Social Sciences*, Seminar Press: New York.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12.

Holland, P.W. (1988). Causal inference, path analysis, and recursive structural equations models. in *Sociological Methodology*, Ed C. Clogg, 449–484. American Sociological Association: Washington, DC.

Imbens, G.W. & Angrist, J.D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–476.

Lauritzen, S.L., & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems. *Proceedings of the Royal Statistical Society. Series B* **50**, 154–227.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N., & Leimer, H.G. (1990). Independence properties of directed Markov fields. *Networks* **20**, 491–505.

Manski, C.F. (1990). Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings* **80**, 319–323.

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

Pearl, J. (1993a). Belief networks revisited. *Artificial Intelligence* **59**, 49–56.

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

Pearl, J. (1994a). From bayesian networks to causal networks. In *Bayesian Networks and Probabilistic Reasoning*, Ed A. Gammerman, 1–31. Alfred Walter Ltd.: London.

Pearl, J. (1994b). A probabilistic calculus of actions. In *Uncertainty in Artificial Intelligence*, Eds R. Lopez de Mantaras and D. Poole, 454–462. Morgan Kaufmann: San Mateo, CA.

Pearl, J. (1995). Causal inference from indirect experiments. *Artificial Intelligence in Medicine* **7**, 561–582.

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

Pratt, J.W., & Schlaifer, R. (1988). On the interpretation and observation of laws. *Journal of Econometrics* **39**, 23–52.

Robins, J.M. (1986). 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* **7**, 1393–1512.

Robins, J.M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies.

In *Health Service Research Methodology: A Focus on AIDS*, Eds L. Sechrest, H. Freeman, and A. Mulley, 113–159. NCHSR, U.S. Public Health Service.

Robins, J.M., Blevins, D., Ritter, G., & Wulfsohn, M. (1992). *G*-Estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3**, 319–336.

Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* (General) **147**, 656–666.

Rosenbaum, P., & Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

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

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **7**, 34–58.

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

Simon, H.A. (1953). Causal Ordering and Identifiability. *Studies in Econometric Method*, Eds W.C. Hood and T.C. Koopmans, Chapter 3. John Wiley and Sons: New York.

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

Spiegelhalter, D.J., Lauritzen, S.L., Dawid, P.A., & Cowell, R.G. (1993). Bayesian analysis in expert systems. *Statistical Science* **8**, 219–247.

Spirtes, P. (1994). Conditional independence in directed cyclic graphical models for feedback. To appear in *Networks*.

Spirtes, P., Glymour, C., & Schienes, R. (1993). *Causation, Prediction, and Search.* Springer-Verlag: New York.

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

Wainer, H. (1989). Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. *Journal of Educational Statistics* **14**, 121–140.

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

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* John Wiley and Sons: Chichester, England.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20**, 557–585.