ON THE IDENTIFICATION OF NONPARAMETRIC STRUCTURAL MODELS

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

Abstract

In this paper we study the identification of nonparametric models, that is, models in which both the functional forms of the equations and the probability distributions of the disturbances remain unspecified. Identifiability in such models does not mean uniqueness of structural parameters but rather uniqueness of policy-related predictions that such parameters would normally support.

We provide sufficient and necessary conditions for identifying predictions of the type "Find the distribution of Y, assuming that X is controlled by external intervention," where Y and X are arbitrary variables of interest. Whenever identifiable, such predictions can be expressed in closed algebraic form, in terms of observed distributions. We also show how the identifying criteria can be tested qualitatively using the graphical representation of the structural model, thus simplifying and generalizing the standard identifiability tests of linear models (e.g., rank and order). Finally, we provide meaningful and precise definitions of effect decomposition for both parametric and nonparametric models.

1 Introduction

In the literature on structural equation models, one usually asks whether or not certain or all of the model parameters are identified, that is, does $P(y; \theta_1) = P(y; \theta_2)$ imply $\theta_1 = \theta_2$?[1] Implicit in such questions is the premise that a model is more useful when its parameters are identified. This premise, coupled with the fact that parameter identification plays such a central role in modeling, indicates that analysts attribute to parameters an important metaprobabilistic meaning, one that cannot be expressed in distribution functions.

Indeed, if we adopt an orthodox statistical attitude[2] and pretend that the sole purpose of models is to provide a compact representation for distribution functions, then we should pay no attention to questions of identification. After all, if two distinct and equally parsimonious models are observationally equivalent (i.e., $P(y; \theta_1) = P(y; \theta_2)$ while $\theta_1 \neq \theta_2$) then they should yield the same statistical predictions and any one of the models can be taken as a working hypothesis; whether the model chosen is unique need not concern us. Most modelers, however, are driven by the understanding that the choice of parameters is not arbitrary but has empirical implications that lie outside the distribution functions. To such modelers, the issue of identifiability becomes one of crucial importance.

What are the empirical implications that give the parameters their distinct metaprobabilistic meaning and under what circumstances are those implications uncovered? The standard literature on simultaneous equation models is remarkably vague on this issue. The meaning of the parameters is sometimes described informally as "telling us how data were generated" and sometimes as "measuring the average change in the response variable per unit change in the explanatory variable." Lacking formal definitions for the notions of "generate" and "change" leaves the meaning of the parameters ambiguous and is largely responsible for the long-standing confusion between structural parameters (e.g., path coefficients) and statistical parameters (e.g., regression coefficients).

The purpose of this paper is threefold. First, we will provide a formal definition for the empirical content of structural parameters in terms of control queries, that is, queries about the outcomes of hypothetical controlled experiments. Second, we will extend the notion of identifiability from parametric to nonparametric models by requiring that answers to such control queries, rather than the parameters per se, be identified uniquely from the observed data. In this way, we capture the intent and the ultimate purpose of parameter identification without ever dealing with parameters. Finally, we will devise mathematical procedures for testing whether identifiability holds in a given nonparametric model and show that, in many cases, identifiability can be tested by inspection of the topological features of the diagram associated with the model.

Before moving on to the formal part of the paper (Section 2), it seems appropriate to illustrate the agenda using a simple example.

[1]The term "simultaneous equations" is often used in the literature, interchangeably with "structural equations". We prefer the latter and will define precisely what makes a set of equations "structural" (see Subsections 1.2 and 2.1). We will also take the liberty of using the symbol $P(\cdot)$ to denote the probability functions for both continuous and discrete variables.

[2]Unfortunately, most statisticians and some social scientists still adhere to this attitude.

1.1 Parametric vs. nonparametric models, an example

Consider the following set of structural equations:

$$
\begin{aligned}
X &= f_1(U) && (1)\\
Z &= f_2(X, V) && (2)\\
Y &= f_3(Z, U, W) && (3)
\end{aligned}
$$

where X, Z, and Y are observed variables, f_1, f_2, and f_3 are unknown arbitrary functions, and U, V, and W are unobservables that we can regard either as latent variables or as disturbances. For the sake of this discussion, we will assume that U, V, and W are mutually independent and arbitrarily distributed. Such a set of equations would obtain, for example, when the modeler is not willing to commit to any particular functional form but feels strongly about the qualitative nature of the data-generating process – for example, that X is a factor determining Z, that X and Y are influenced by some common factor U, and so on. Graphically, these influences can be represented by the path diagram of Figure 1. Note that the arcs in Figure 1 should be labeled with the functions themselves, not with coefficients as in traditional path analysis where all relationships are assumed linear.

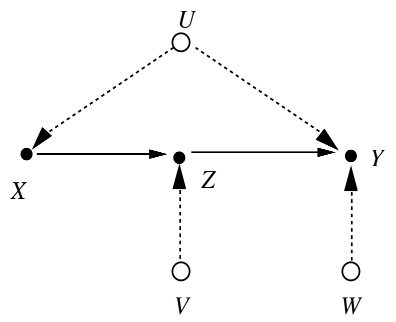

Figure 1:
Path diagram corresponding to Eqs. (1)-(3), where $\{X, Z, Y\}$ are observed and $\{U, V, W\}$ are unobserved.

The problem we pose is as follows:
We have drawn a long stream of independent samples of the process defined by Eqs. (1)-(3) and have recorded the values of the observed quantities X, Z, and Y, and we now wish to estimate the unspecified quantities of the model to the greatest extent possible.

To sharpen the scope of our problem, let us consider its solution in the case of a linear version of the model, which is given by[3]

$$
\begin{aligned}
X &= U && (4)\\
Z &= aX + V && (5)\\
Y &= bZ + cU + W && (6)
\end{aligned}
$$

[3]An equivalent version of this model would obtain by eliminating cU from the equation of Y and allowing U and W to be correlated.

with $\{U, V, W\}$ uncorrelated, zero-mean disturbances. It is not hard to show that all three parameters, a, b and c, can be determined uniquely from the correlations among the observed quantities X, Z and Y. In particular, multiplying Eq. (5) by X and taking expectations gives

$$a = \frac{E(XZ)}{E(X^2)} \tag{7}$$

Further multiplying Eq. (6) by X, then by Z, taking expectations and solving, gives

$$b = \frac{E(XZ)\ E(XY) - E(X^2)\ E(ZY)}{E^2(XZ) - E(Z^2)\ E(X^2)} \tag{8}$$

$$c = \frac{E(XY)\ E(Z^2) - E(ZY)E(ZX)}{E(X^2)\ E(Z^2) - E^2(ZX)} \tag{9}$$

Thus we see that, given the right set of assumptions on the disturbances, together with the parametric form of the equations, it is possible to estimate all model parameters from the observed distribution; such models are called "identifiable."

Returning to our nonparametric model of Eqs. (1)-(3), a natural generalization would be to require that, for the model to be identifiable, the functions $\{f_1, f_2, f_3\}$ be determined uniquely from the data. However, this prospect is doomed to failure from the start. When the equations are in nonparametric form, we are generally unable to identify the form of the functions, even when the disturbance distributions are known precisely. Consider the simplest case of an equation containing just two variables, for example, $Z = f_2(X,\ V)$. We are unable to determine the functional form of f_2 from the joint distribution $P(x, z)$ of X and Z, even if we are given the distribution $P(v)$ and know that V is independent of X. In other words, there are many functions f_2 compatible with the given distributions, each defining a different mapping from $\{X,\ V\}$ to Z. Thus, it might appear that the problem of identifying nonparametric models is hopeless and that nothing useful can be inferred from such loosely specified model as the one given in Eqs. (1)-(3).

However, parametric and functional identification is not an end in itself, even in linear models, but rather serves to answer practical questions of prediction and control. Therefore, the right question to ask is not whether the data permit us to identify the form of the equations but rather whether the data can constrain the equations to the extent of providing unambiguous answers to questions of interest, of the kind answered by traditional parametric models.

If the model is used merely for probabilistic prediction (i.e., to determine the probabilities of some variables given a set of observations on other variables), then such predictions can be estimated directly from the observed distributions. Moreover, if dimensionality reduction is needed (e.g., to improve estimation accuracy) these distributions can be encoded in a variety of simultaneous equation models, all of the same dimensionality. For example, the correlations among X, Y and Z in the linear model, M, of Eqs. (4)-(6) might as well be represented by the model M':

$$X = U \tag{10}$$

$$Z = a'X + V \tag{11}$$

$$Y = b'Z + dX + W \tag{12}$$

which is as compact as Eqs. (4)-(6). Although obviously the choice is not unique, it is nevertheless compatible with the observations and, upon setting $a' = a, b' = b$ and $d = c$, will yield the same probabilistic predictions as would the model of Eqs. (4)-(6) . Still, when viewed as data-generating mechanisms, the two models are clearly not equivalent; each tells a different story of the processes generating X, Y and Z and, naturally, each predicts different changes that would result from subjecting these processes to external interventions.

1.2 Causal effects: The structural interpretation of simultaneous equation models

The difference between the two models above illustrates precisely where the structural reading of simultaneous equation models comes into play.[4] Model M', defined by Eqs. (10)-(12), proclaims X to be a direct participant in the process which determines the value of Y, while model M, defined by Eqs. (4)-(6), views X as an indirect factor; its effect on Y is mediated by Z. This difference is not manifested in the data but only in the way the data would change in response to outside interventions. For example, suppose we wish to predict the expectation of Y after we intervene and fix the value of X to some constant x, denoted $E(Y|set(X = x))$. Substituting $X = x$ in Eq. (11) and (12), model M' yields

$$\begin{aligned} E[Y|set(X = x)] &= E[b'a'x + b'V + dx + W] && (13) \\ &= (b'a' + d)x && (14) \end{aligned}$$

while model M yields

$$\begin{aligned} E[Y|set(X = x)] &= E[bax + bV + cU + W] && (15) \\ &= bax && (16) \end{aligned}$$

Equating $a' = a$, $b' = b$ and $d = c$ (to match the data, as in Eqs. (7)-(9)), we see clearly that the two models assign different magnitudes to the (total) effect of X on Y; model M predicts that a unit change in x will change $E(Y)$ by an amount ba, while model M' puts this amount at $ba + c$.

At this point, it is natural to ask whether we should not substitute the constant x for U in Eq. (6) prior to taking expectations in Eq. (15). If we permit the substitution of Eq. (5) in Eq. (6), so the argument goes, why not substitute Eq. (4) as well? After all, there is no harm in upholding a mathematical equality, $U = X = x$, which the modeler deems valid. This argument is fallacious. Structural equations are not meant to be treated as immutable mathematical equalities. Rather, they are introduced into the model to describe an equilibrium condition, only to be violated when that equilibrium is perturbed by outside interventions. The power of structural models is that they also encode the

[4] I ask the reader to bear with me as I review concepts that might seem obvious. The reviewer of this paper has commented: "Nor is it clear to me that structural equations do anything more than summarize distributions" and has proposed specifically that the model in Eqs. (4)-(6) be replaced with that of Eqs. (10)-(12). To me, these comments suggest that even renowned scholars and experienced modelers do not find the interpretation of structural equations to be as obvious as one would expect; neither have I found these issues addressed forthrightly in the standard literature. Subsection 3.4 discusses some of the prevailing confusions and the reasons that they have not been resolved thus far.

information necessary for determining the new equilibrium. If the intervention consists merely of holding X constant, at x, then the equation $X = U$, which represents the pre-intervention process determining X, should be overruled and replaced with the equation $X = x$. The solution to the new set of equations then represents the new equilibrium. Thus, the essential characteristic of structural equations, which sets them apart from ordinary mathematical equations is that they do not stand for one, but for many sets of equations, each corresponding to a subset of equations taken from the original model. Every such subset represents some physical reality, one that is configured by overruling all but the processes corresponding to the selected equations.

Taking the stand that the primary value of structural equations lies not in summarizing distribution functions but in encoding causal information for predicting the effect of interventions [Haavelmo 1943, Marschak 1953], it is natural to view such predictions as the proper generalization of structural coefficients when dealing with nonparametric model. For example, the proper generalization of the coefficient b in the linear model M would be the answer to the control query: "What would be the change in the expected value of Y if we were to intervene and change the value of Z from z to $z+1$," which is different, of course, from the observational query "What would be the difference in the expected value of Y if we were to find Z at level $z+1$, instead of z." Observational queries can be answered directly from the joint distribution $P(x, y, z)$, while control queries require causal information, such as the one encoded in structural equations, as well. To distinguish between the two types of queries, we use the "hat" symbol ($\hat{}$) to indicate externally controlled quantities. For example, we write

$$E(Y|\hat{x}) = E[Y|set(X = x)] \tag{17}$$

for the controlled expectation and

$$E(Y|x) = E(Y|X = x) \tag{18}$$

for the standard conditional expectation. The inequality $E(Y|\hat{x}) \neq E(Y|x)$ can easily be seen in the model of Eqs. (4)-(6), where $E(Y|\hat{x}) = abx$ while $E(Y|x) = (ab + c)x$. Indeed, the passive observation of $X = x$ should not violate any of the equations, and would justify substituting Eqs. (4) and (5) in (6) before taking the expectation.

In the case of linear models, the answers to questions of direct control are encoded in the so-called "path coefficients" or "structural coefficients," and these can be used to derive the total effect of any variable on another. For example, the value of $E(Y|\hat{x})$ in the model defined by Eqs. (4)-(6) is abx, i.e., x times the product of path coefficients along the path $X \longrightarrow Z \longrightarrow Y$. In the nonparametric case, the computation of $E(Y|\hat{x})$ would naturally be more complicated, even when we know the functions f_1, f_2, and f_3. It is nevertheless well defined, and requires the solution (for the expectation of Y) of a modified set of equations in which f_1 is "wiped out" and X is replaced by the constant x:

$$\begin{aligned} Z &= f_2(x, V) && (19) \\ Y &= f_3(Z, U, W) && (20) \end{aligned}$$

Thus, the computation of $E(Y|\hat{x})$ requires the evaluation of

$$E(Y|\hat{x}) = E\{f_3[f_2(x, V), U, W]\}$$

where the expectation is taken over U, V, and W. This computation will be carried out in Section (2.3). Similar modifications of the model are required for the computation of $E(Z|\hat{x})$, $E(X|\hat{z})$, or $E(X|\hat{y})$, and can easily be shown to yield $E(Z|\hat{x}) = E(Z|x)$, $E(X|\hat{z}) = E(X)$, and $E(X|\hat{y}) = E(X)$, respectively.

What then would be an appropriate definition of "identifiability" for nonparametric models? Consistent with our focus on control queries, a reasonable definition of identifiability is that answers to such queries are *unique.* Accordingly, we will define a model to be *identifiable* if there exists a consistent estimate for every control query of the type "Find $P(r|\hat{s}) = P[R = r|set(S = s)]$," where R and S are subsets of observables and r and s are any realization of these variables. The set of probabilities $P(r|\hat{s})$ is called the "causal effect" of S on R, as it describes how the distribution of R varies when S is changed by external control.[5] Naturally, we should allow for some queries to be identifiable while the system as a whole is not. Hence, we say that $P(r|\hat{s})$ is identifiable in model M if every choice of model parameters (i.e., the functional forms and the distributions) that is compatible with the observed distribution P would yield the same value for $P(r|\hat{s})$.

For example, we might inquire whether the model defined by Eqs. (1)-(3) is identifiable. The answer is yes; we will see that this model permits the identification of all control queries. For example, the methods developed in Section 2 will enable one to conclude immediately that:

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

2. $P(z|\hat{x}) = P(z|x)$,
   because V is independent of X, hence Z is not confounded with X; and

3. $P(y|\hat{z}) = \sum_x P(y|z, x)P(x)$,
   because x is an appropriate covariate for adjustment; and

4. $P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x')$,
   for reasons to be explained in Section 2.

These answers are unique because all terms on the right-hand sides are functions of the observable distribution $P(x, y, z)$. Hence, any choice of functions (f_1, f_2, and f_3) and distributions (of U, V, and W) compatible with the observed distribution P would necessarily yield the same answers to the control queries above.

Remarkably, many aspects of nonparametric identification, including tests for deciding whether a given control query is identifiable, as well as formulas for estimating such queries, can be determined graphically, almost by inspection, from the path diagram. These aspects will be developed and demonstrated in the body of the paper.

[5]Technically, the adjective "causal" is redundant. It merely serves to emphasize, however, that the changes in S are enforced by external control, and do not represent stochastic variations in the observed value of S. The phrase "the effect of S on R" has improperly been applied to $P(r|s)$, in which s stands for uncontrolled statistical observations.

2 Computing Causal Effects

2.1 Definitions and Notation

2.1.1 Models, Graphs, and Theories

We consider models consisting of a set of n (recursive) equations

$$X_i = f_i(X_1, X_2, ..., X_{i-1}; U_1, ..., U_m),\ \ i = 1, 2, ..., n, \tag{21}$$

where $X_1, ..., X_n$ are observed variables, and $U_1, ..., U_m$ are unobserved (or latent) disturbances.[6] The f_i are unspecified deterministic functions with restricted sets of arguments, and the distribution of the disturbances may be constrained by independence restrictions but is otherwise unspecified.

Restrictions on the arguments of the equations[7] can be represented by a directed graph G in which each node corresponds to an observed variable and an arrow from node X_i to node X_j indicates that X_i is an argument of f_j. The restrictions on the dependencies among the U variables will also be represented graphically, by adding a "confounding path" (a dashed curved arc with double arrows) between any two variables X_i and X_j whenever a dependency exists between the U variables in f_i and those in f_j. Thus, for example, the model described by Eqs. (1)-(3) is completely specified by the graph of Figure 2. Unlike the path diagram of Figure 1, G does not represent the disturbances

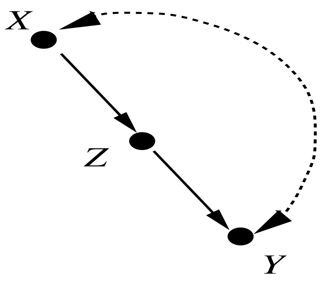

Figure 2:
A graph G, representing the restrictions specified in the nonparametric model of Eqs. (1)-(3).

explicitly; only the dependencies induced by these disturbances are represented. Each confounding path may represent several unobserved disturbances common to a given pair of equations. In most cases, modelers prefer to specify the induced dependencies directly without making the disturbances explicit. However, in order to read off the dependencies embodied in a given graph, it sometimes may be convenient to restore the U variables. In

[6] The recursive nature of Eq. (21) corresponds to a lower triangular matrix B in linear models [Bollen 1989] and therefore excludes feedback mechanisms. This restriction is not essential to the discussion of the basic concepts, though it simplifies the test for identifiability. The restriction that all unobserved variables be exogenous (i.e., do not appear on the left-hand side of any equation) can easily be relaxed: endogenous latent variables can always be eliminated by substitutions, thus restoring the form of Eq. (21).

[7] In linear models, these correspond to the "zero-coefficient" restrictions, while independencies among the U's are specified by zero entries in the covariance matrix of the disturbances.

such cases, we will add a "dummy" root node (hollow circle) for each dashed arc, as shown by the node marked U in Figure 1. A theorem by [Pearl & Verma 1991] states that such "dummy" root nodes can faithfully represent any pattern of dependencies among any set of latent variables.

The structural model M defined by Eq. (21) (equivalently, by the corresponding graph $G(M)$) delineates a set of *grounded models* which we call *theories.*[8] Each theory $T =< \{f_i\},\ P(u) >$ in M corresponds to a specific choice of function f_i and a specific choice of disturbance distribution $P(u) = P(u_1, ..., u_m)$, both satisfying the restrictions imposed by M. For each theory T of M, there is a corresponding unique probability distribution $P_T(x) = P_T(x_1, ..., x_n)$, which we say to be "generated" by T.

Definition 1 *We say that $P(x)$ is* compatible *with M iff there exists a theory T of M that generates $P(x)$, i.e.,*

$$P(x) = P_T(x)$$

A model M is said to be universal *if it is compatible with every arbitrary $P(x)$; otherwise, it is said to be* falsifiable. □

Clearly, every model whose corresponding graph is complete (i.e., every pair is connected by an arrow) is universal, since such a model can generate any given $P(x)$, using mutually independent disturbances. Figure 2 is an example of a universal model which will become falsifiable upon removing any of the arcs.

2.1.2 Queries and Identifiability

A query q is any quantity that can be computed from a given theory; i.e., a functional of T. For example, the queries

$$\begin{aligned} q_1 : &\quad f_2(X = 1, V = 3.06) = ? \\ q_2 : &\quad P(U = 1|Y = 0.8) = ? \\ q_3 : &\quad P(X = 1|Y = 3) = ? \\ q_4 : &\quad E_u[P(Y = 1|X = 1, u) = ? \end{aligned} \tag{22}$$

can be computed from any theory of the structural model described in Eqs. (1)-(3), because, once we choose the functions $\{f_1,\ f_2,\ f_3\}$ and the distribution $P(u, v, w)$, the answers to each of these queries is well defined. Note, moreover, that the answers to queries $q1$ and $q2$ depend critically on the specific choice of theory, while $q3$ depends solely on the distribution $P(x, y, z)$. Query $q4$, which the reader may recognize as $P(Y = 1|set(X = 1))$ (Section 1), appears at first glance to depend on the choice of theory. We will see, however, that, by virtue of the structural restrictions communicated by Figure 2, query q_4 will have the same answer in all theories that generate a given distribution $P(x, y, z)$. This motivates the following definition of identifiability:

Definition 2 *(identifiability) A query q is said to be* identifiable *in a model M iff, for any two theories T_1 and T_2 of M,*

$$q(T_1) = q(T_2),\ \text{whenever}\ P_{T_1}(x) = P_{T_2}(x)\ \text{and}\ P_{T_1}(x) > 0$$

[8]Koopmann and Reiersol (1950) used the term "structure" for our "theory."

A model M is said to be identifiable relative to a set Q of queries if every member of Q is identifiable in M. □

In other words, a model M is identifiable relative to a set Q of queries if every query in Q can be computed uniquely from the pair $\{M, P\}$, where P is any positive distribution over the observables that is compatible with M.

Technically, the reason for restricting the observed distributions to positive distributions is to avoid conditioning on events with zero probabilities. Conceptually, positivity ensures that each function is perturbed by some stochastic disturbance; these disturbances act like instrumental variables, or randomized experiments conducted by nature, in that they help reveal the strength of causal effects of some variables while others are kept constant.

It is clear, from the definition above, that every model is trivially identifiable relative to queries, such as q_3, which are addressed to the observed distribution $P(x)$. As discussed in the introduction, the focus of this paper is the set of *control queries*, like q_4, that are the primary (yet often forgotten) reason we use structural modeling [Haavelmo 1943].

2.1.3 Control Queries

Of special interest to us will be the set Q_2 of *pairwise* control queries, in which each query is of the type "Find the distribution of X_j given that X_i is held fixed at x_i", where i and j are arbitrary. Answers to such queries are the nonparametric analogs to the so-called "causal effects" or "total effects" in linear models, and in these models the answers can be computed directly from the structural coefficients. To answer such queries, we need to formalize the notion of "holding fixed" within the general framework of nonparametric structural models.

Given a model M and a subset S of variables, define a submodel M_s of M as the set of equations that results if the $|S|$ equations corresponding to the variables in S are deleted from M and $S = s$ is substituted in the remaining equations. For example, the model specified in Eqs. (19)-(20) is the submodel M_x of the model in Eqs. (1)-(3). The theories delineated by M_s will be denoted by T_s.

Definition 3 *(control queries) A control query $q = P(r|\hat{s})$ (read: the probability of $R = r$ given that S is* held fixed *at s) is a functional of the theories of M, defined by*

$$P_T(r|\hat{s}) = P_{T_s}(r).$$

In other words, the value of $P(r|\hat{s})$ in theory T is given by the probability $P(r)$ induced by the subtheory T_s of T. □

The notion of subtheories reflects the understanding that external interventions perturb the normal causal influences as represented by the structural equations. In particular, the primitive intervention "holding X fixed" has sharp, local effect on those mechanisms, that is, it totally neutralizes X from its normal influences and places it under a new influence (given by the intervention), while keeping all other influences unperturbed. This interpretation of control queries is an integral part of viewing structural equations as representing a set of *autonomous*, *stable*, or *invariant* mechanisms—a notion going back to [Frisch 1938, Haavelmo 1943, Marschak 1953] and later expanded by [Simon 1977] and [Goldberger 1973].[9] 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] for defining effects decomposition (see Section 3.4). Formal graphical accounts of this notion are given in [Spirtes et al. 1993] and [Pearl 1993].

2.1.4 Graphs, Conditional Independence, and d-Separation

In this subsection, we review the properties of directed acyclic graphs (DAGs) as carriers of conditional independence information [Pearl 1988]. Readers familiar with this aspect of DAGs are advised to skip to Section 2.2.

Given a DAG G and a joint distribution P over a set $V = \{X_1, ..., X_n\}$ of variables, we say that G *represents* P if there is a one-to-one correspondence between the variables in X and the nodes of G, such that P admits the product decomposition

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid pa_i) \tag{23}$$

where pa_i are the values of the direct predecessors (called *parents*), PA_i, of X_i in G. For example, the DAG in Figure 3 induces the decomposition

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1)\ P(x_4|x_2, x_3)\ P(x_5|x_4) \tag{24}$$

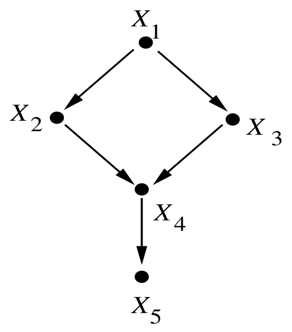

Figure 3:
A typical directed acyclic graph (DAG) representing the decomposition of Eq. (24).

A convenient way of characterizing the set of distributions represented by a DAG G is to list the set of (conditional) independencies that each such distribution must satisfy. Clearly, the decomposition in Eq. (23) implies (using the chain rule) that, given its parent set PA_i, each variable X_i is conditionally independent of all its other predecessors

[9] As discussed briefly in Section 1, the notion of invariance, and its operational derivative of Definition 3 is, in fact, the defining feature of structural equations. We, therefore, depart from the views of [Sobel 1990] and others and do not refer to invariance as an assumption that requires further justification or judgment. In other words, invariance is what the investigator must already have in mind when he/she specifies the arguments of each function f_i in the model.

$\{X_1, X_2, ..., X_{i-1}\} \backslash PA_i$. We call this set of independencies *Markovian*, because it reflects the Markovian condition for state transitions: Each state is rendered independent of the past, given its immediately preceding state. However, the decomposition of Eq. (23) implies additional, less obvious independencies which can be read off the DAG by using a graphical criterion called d-separation [Pearl 1988]. To test whether X is independent of Y given Z in the distributions represented by G, we need to examine G and test whether the nodes corresponding to variables Z d-separate all paths from nodes in X to nodes in Y. By *path* we mean a sequence of consecutive edges (of any directionality) in the DAG.

Definition 4 (d-separation) *A path p is said to be d-separated (or blocked) by a set of nodes Z iff:*

(i) p contains a chain $i \longrightarrow j \longrightarrow k$ or a $fork$ $i \longleftarrow j \longrightarrow k$ such that the middle node j is in Z, or,

(ii) p contains an inverted fork $i \longrightarrow j \longleftarrow k$ such that neither the middle node j nor any of its descendants (in G) are in Z.

If X, Y, and Z are three disjoint subsets of nodes in a DAG G, then Z is said to d-separate X from Y, denoted $(X \perp\!\!\!\perp Y)_G$, iff Z d-separates every path from a node in X to a node in Y.

The intuition behind d-separation is simple: In chains $X \to Z \to Y$ and forks $X \leftarrow Z \to Y$, the two extreme variables are dependent (marginally) but become independent of each other (i.e., blocked) once we know the middle variable. Inverted forks $X \to Z \leftarrow Y$ act the opposite way; the two extreme variables are independent (marginally) and become dependent (i.e., unblocked) once the value of the middle variable (i.e., the common effect) or any of its descendants is known. For example, finding that the pavement is wet or slippery (see Figure 1) renders Rain and Sprinkler dependent, because refuting one of these explanations increases the probability of the other.

In Figure 1, for example, $X = \{X_2\}$ and $Y = \{X_3\}$ are d-separated by $Z = \{X_1\}$; the path $X_2 \leftarrow X_1 \to X_3$ is blocked by $X_1 \in Z$, while the path $X_2 \to X_4 \leftarrow X_3$ is blocked because X_4 and all its descendants are outside Z. Thus $(X_2 \perp\!\!\!\perp X_3 | X_1)_G$ holds in G. However, X and Y are not d-separated by $Z' = \{X_1, X_5\}$, because the path $X_2 \to X_4 \leftarrow X_3$ is unblocked by virtue of X_5, a descendant of X_4, being in Z'. Consequently, $(X_2 \perp\!\!\!\perp X_3 | \{X_1, X_5\})_G$ does not hold; in words, learning the value of the consequence X_5 renders its causes X_2 and X_3 dependent, as if a pathway were opened along the arrows converging at X_4.

Theorem 1 [Verma & Pearl 1990, Geiger et al. 1990]. *For any three disjoint subsets of nodes (X, Y, Z) in a DAG G, Z d-separates X from Y in G implies that X is independent of Y conditional on Z in every probability distribution represented by G.*

Thus, a DAG can be viewed as an efficient scheme for representing Markovian independence assumptions and for deducing and displaying all the logical consequences of such assumptions. Note that the precise ordering of the nodes does not enter into the

d-separation criterion; it is only the topology of the graph that determines the set of independencies that the probability P must satisfy.

An important property that follows from the d-separation characterization is a criterion for determining whether two given DAGs are observationally equivalent, that is, whether every probability distribution that is represented by one of the DAGs is also represented by the other.

Theorem 2 [Verma & Pearl 1990] *Two DAGs are observationally equivalent iff they have the same sets of edges and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow.*

Observational equivalence places a limit on our ability to infer the directionality of the links directionality from probabilities alone. For example, reversing the direction of the arrow between X_1 and X_2 in Figure 1 does not introduce any new v-structure. Therefore, this reversal yields an observationally equivalent DAGs, and the directionality of the link $X_1 \rightarrow X_2$ cannot be determined from probabilistic information. The arrows $X_2 \rightarrow X_4$ and $X_4 \rightarrow X_5$, however, are of different nature; there is no way of reversing their directionality without creating a new v-structure. Thus, we see that some probability functions P can constrain the directionality of some arrows in their DAG representation.

Additional properties of DAGs and their applications to evidential reasoning are discussed in [Geiger 1990, Lauritzen & Spiegelhalter 1988, Spiegelhalter et al. 1993, Pearl 1988, Pearl 1993, Pearl et al. 1990].

2.2 A Causal Calculus

This subsection establishes a set of sound (and possibly complete) inference rules by which probabilistic sentences involving actions and observations can be transformed to other such sentences, thus providing a syntactic method for deriving (or verifying) claims about actions and observations. Given the pair $< M, P >$, our main problem will be to facilitate the syntactic derivation of expressions of the form $P(x_j|set(x_i))$ from standard probability expressions.

Let X, Y, and Z be arbitrary disjoint sets of nodes in a DAG G. We denote by $(X \perp\!\!\!\perp Y|Z)_G$, the proposition "$Z$ d-separates X from Y in G" (see Definition 4). We denote by $G_{\overline{X}}$ ($G_{\underline{X}}$, respectively) the graph obtained by deleting from G all arrows pointing to (emerging from, respectively) nodes in X. In dealing with expressions involving both observed and fixed variables, we will use $P(y|\hat{x}, z)$, where the $\hat{}$ symbol identifies the variables that are kept constant externally. In words, the expression $P(y|\hat{x}, z)$ will stand for the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x.

Armed with this notation, we formulate the three basic inference rules of our calculus in the following theorem [Pearl 1995a]:

Theorem 3 *Let G be a DAG characterizing a structural model M, and let P be a distribution generated by some theory of M. Then, for any disjoint sets of variables X, Y, Z, and W, we have:*

Rule 1 *Insertion/deletion of observations*

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{25}$$

Rule 2 *Action/observation exchange*

$$P(y|\hat{x},\hat{z},w) = P(y|\hat{x},z,w) \quad if \quad (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\underline{Z}}} \tag{26}$$

Rule 3 *Insertion/deletion of actions*

$$P(y|\hat{x},\hat{z},w) = P(y|\hat{x},w) \quad if \quad (Y \perp\!\!\!\perp Z|X,\ W)_{G_{\overline{X},\ \overline{Z(W)}}} \tag{27}$$

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 can be proven [Pearl 1995a] from the basic interpretation of the "set(x)" operation as a replacement of the causal mechanism that connects X to its parents prior to the operation with a new mechanism $X = x$ introduced by the intervention (Definition 3). This results in a submodel M_x which is characterized by the subgraph $G_{\overline{X}}$ (named "manipulated graph" in [Spirtes et al. 1993]).

Rule 1 reaffirms *d*-separation as a valid test for Bayesian conditional independence in the distribution determined by the intervention $set(X = x)$, hence the graph $G_{\overline{X}}$.

Rule 2 provides conditions 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 2 is equivalent to the "back-door criterion"[10] of [Pearl 1993] and can also be derived from Theorem 7.1 in [Spirtes et al. 1993].

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 severing all relations between Z and its parents (hence the graph $G_{\overline{XZ}}$).

Corollary 1 *A query q: $P(y_1, ..., y_k|\hat{x}_1, ..., \hat{x}_m)$ is identifiable in model M if there exists a sequence of inference rules which transforms q into a standard (i.e., hat-free) probability expression.*

2.3 Computing Causal Effects: An Example

We will now demonstrate how these inference rules can be used to evaluate all control queries for the structural model specified in Eqs. (1)-(3). The graphical characterization

[10] The back-door criterion states that, if there exists a set S of observed variables which are nondescendants of X and which block every back-door path from X to Y (that is, paths ending with arrows pointing to X), then $P(y|\hat{x})$ is identifiable and is given by the formula

$$P(y|\hat{x}) = \sum_s P(y|x,s)P(s) \tag{28}$$

of this model is given by the DAG G of Figure 4, which is identical to that of Figure 2 save for the explicit representation of the unobserved variable U. We will see that this structure permits us to quantify, using the causal calculus of Section 2.2, the effect of every action on every set of observed variables. Our task amounts to reducing expressions involving actions to those involving only observations, that is, to eliminating the "hat" symbol ($\hat{}$) from the query expressions.

The applicability of the inference rules in Theorem 3 requires that the d-separation conditions holds in various subgraphs of G, and the structure of each subgraph varies with the expressions to be manipulated. Figure 4 displays the subgraphs that will be needed for the derivations that follow.

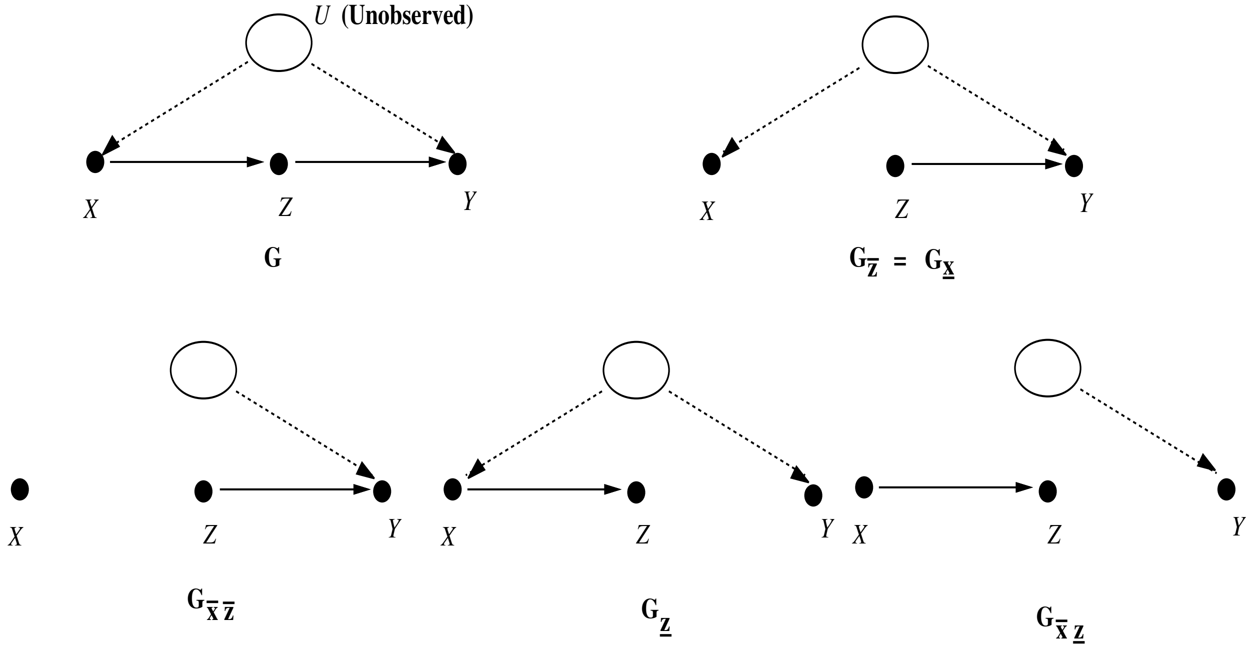

Figure 4:
Subgraphs of G used in the derivation of causal effects.

Task-1, compute $P(z|\hat{x})$
This task can be accomplished in one step, since G satisfies the applicability condition 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 collider at Y) and we can write

$$P(z|\hat{x}) = P(z|x) \tag{29}$$

Task-2, compute $P(y|\hat{z})$
Here we cannot apply Rule 2 to replace $\hat{z}$ with z, because $G_{\underline{Z}}$ contains a path from Z to Y (a so-called "back-door" path). Naturally, we would like to "block" this path by "adjusting for" covariates (such as X) that reside on that path. Symbolically, the "adjustment" operation involves conditioning and summing over all values of X, as follows:

$$P(y|\hat{z}) = \sum_x P(y|x, \hat{z}) P(x|\hat{z}) \tag{30}$$

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

$$P(x|\hat{z}) = P(x) \ \text{ if } \ (Z \perp\!\!\!\perp X)_{G_{\overline{Z}}} \tag{31}$$

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

$$P(y|x,\hat{z}) = P(y|x,z) \ \text{ if } \ (Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}} \tag{32}$$

and note that X d-separates Z from Y in $G_{\underline{Z}}$. This allows us to write Eq. (30) as

$$P(y|\hat{z}) = \sum_x P(y|x,z)P(x) = E_x P(y|x,z) \tag{33}$$

which is a special case of the "back-door" formula (see footnote 10, (28)) with $S = X$. This formula appears in a number of treatments of causal effects (e.g., [Rosenbaum & Rubin 1983, Rosenbaum 1989, Pratt & Schlaifer 1988]) in which the legitimizing condition $(Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}}$ is expressed in terms of conditional-independence judgments involving counterfactual variables. The causal calculus facilitated by Theorem 3 replaces such complicated judgments with formal tests (d-separation) on a graph (G) which represents familiar processes.

We are now ready to tackle a harder task—the evaluation of $P(y|\hat{x})$, which cannot be reduced to an observational expression by direct application of any of the inference rules.

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

Writing

$$P(y|\hat{x}) = \sum_z P(y|z,\hat{x})P(z|\hat{x}) \tag{34}$$

we see that the term $P(z|\hat{x})$ was reduced in Eq. (29) but that no rule can be applied to eliminate the manipulation symbol $\hat{}$ from the term $P(y|z,\hat{x})$. However, we can add a $\hat{}$ symbol to this term via Rule 2

$$P(y|z,\hat{x}) = P(y|\hat{z},\hat{x}) \tag{35}$$

since Figure 3 shows

$$(Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}\underline{Z}}}$$

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

$$P(y|z,\hat{x}) = P(y|\hat{z}) \tag{36}$$

which was calculated in Eq. (33). Substituting Eqs. (33), (36), and (29) back into Eq. (34) yields

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x',z)P(x') \tag{37}$$

Eq. (37) was named the "front-door" formula in [Pearl 1995b], as it involves a (nonstandard) adjustment for a variable (Z) that stands between the cause (X) and the effect (Y).

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 right-hand side were derived in Eqs. (29) and (36), from which we obtain

$$\begin{aligned} P(y, z|\hat{x}) &= P(y|\hat{z})P(z|x) \\ &= P(z|x) \textstyle\sum_{x'} P(y|x', z)P(x') \end{aligned} \tag{38}$$

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

$$\begin{aligned} P(x, y|\hat{z}) &= P(y|x, \hat{z})P(x|\hat{z}) \\ &= P(y|x, z)P(x) \end{aligned} \tag{39}$$

The first term on the right is obtained by Rule 2 (licensed by $G_{\underline{Z}}$) and the second term, by Rule 3 (as in Eq. (31)).

3 Graphical Tests of Identifiability

In the example above, 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. A bow-pattern represents an equation

$$Y = f_Y(X, U)$$

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 $b \triangleq \frac{\partial}{\partial x}E(Y|\hat{x})$ 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 b via the instrumental-variable formula [Bowden & Turkington 1984, Bollen 1989]:

$$b \triangleq \frac{\partial}{\partial x}E(Y|\hat{x}) = \frac{E(Y|z)}{E(X|z)} = \frac{R_{yz}}{R_{xz}} \tag{40}$$

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 [Pearl 1995b]. 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 dependence between compliance and response, as is done, for example, by Angrist et al. (1993) and Imbens & Angrist (1994). While the added arc $Z \rightarrow X$ permits us to calculate bounds on $P(y|\hat{x})$ [Robins 1989, Section 1g],[Manski 1990, Balke & Pearl 1994], and the upper and lower bounds may even coincide for certain types of distributions $P(x, y, z)$ [Balke & Pearl 1993, Pearl 1995b] there is no way of computing $P(y|\hat{x})$ for every positive distribution $P(x, y, z)$, as required by Definition 2. It is interesting to note that the noncompliance model of Figure 9(b) is falsifiable whenever X is discrete, but has no testable implications when X is continuous [Pearl 1995c] .

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 1) would succeed in the original diagram.

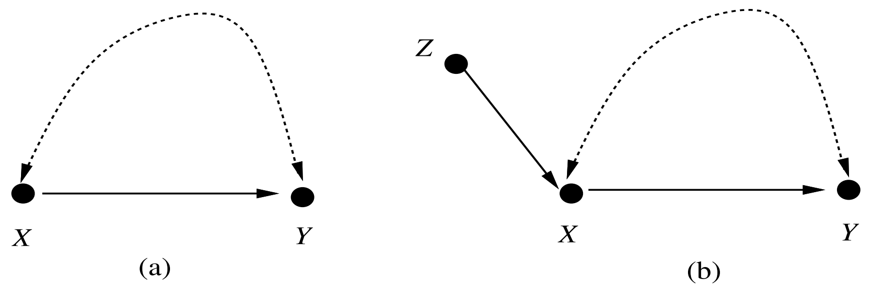

Figure 5:
(a) A bow-pattern: a confounding arc embracing a causal link $X \rightarrow Y$, thus preventing the identification of $P(y|\hat{x})$ even in the presence of an instrumental variable Z, as in (b).

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 6, 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, the graph shown in Figure 6 is the smallest graph that does not contain the bow-pattern of Figure 5 and still presents an uncomputable causal effect.

Another interesting feature demonstrated by Figure 6 is that computing the effect of a joint action is often easier than computing the effects of its constituent singleton actions.[11]

[11] 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. (41) for example, can be obtained from Robin's G-computation formula [Robins 1989, Robins et al. 1992].

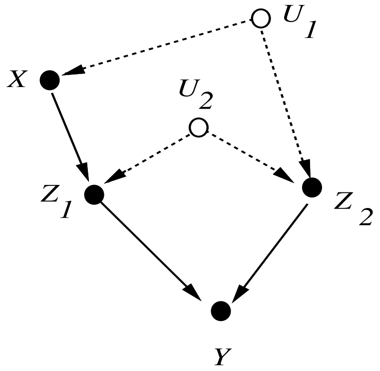

Figure 6:
A graph not containing a bow, but still prohibiting the identification of $P(y|\hat{x})$.

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

$$\begin{aligned} P(y|\hat{x}, \hat{z}_2) &= \sum_{z_1} P(y|z_1, \hat{x}, \hat{z}_2) P(z_1|\hat{x}, \hat{z}_2) \\ &= \sum_{z_1} P(y|z_1, x, z_2) P(z_1|x) \end{aligned} \tag{41}$$

The computation of $P(y|\hat{x})$, on the other hand, requires the conversion of $P(z_1|\hat{x}, z_2)$ into $P(z_1|x, z_2)$; Rule 2 is inapplicable because, when conditioned on Z_2, X and Z_1 are d-connected in $G_{\underline{X}}$ (through the dashed lines). A systematic procedure for identifying causal effects of multiple actions is provided in [Pearl & Robins 1995].

3.1 Identifying Models

Figure 7 shows simple diagrams in which the causal effect of X on Y, $P(y|\hat{x})$, is identifiable. Such structures are called identifying because their structures communicate a sufficient number of assumptions (missing links) to permit the identification of the target quantity $P(y|\hat{x})$. Unobserved (or latent) variables are not shown explicitly in these diagrams; rather, such variables are implicit in the confounding arcs (dashed lines). 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 unobserved variables from the structural equations of Eq. (21) and then constructing a new diagram by connecting any two variables X_i and X_j by (1) an arrow from X_j to X_i whenever X_j appears in the equation for X_i and (2) a confounding arc whenever the same U 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 7.

1. Since the removal of any arc or arrow from a causal diagram can only assist the identifiability of causal effects, $P(y|\hat{x})$ will still be identified in any edge-subgraph of the diagrams shown in Figure 7.

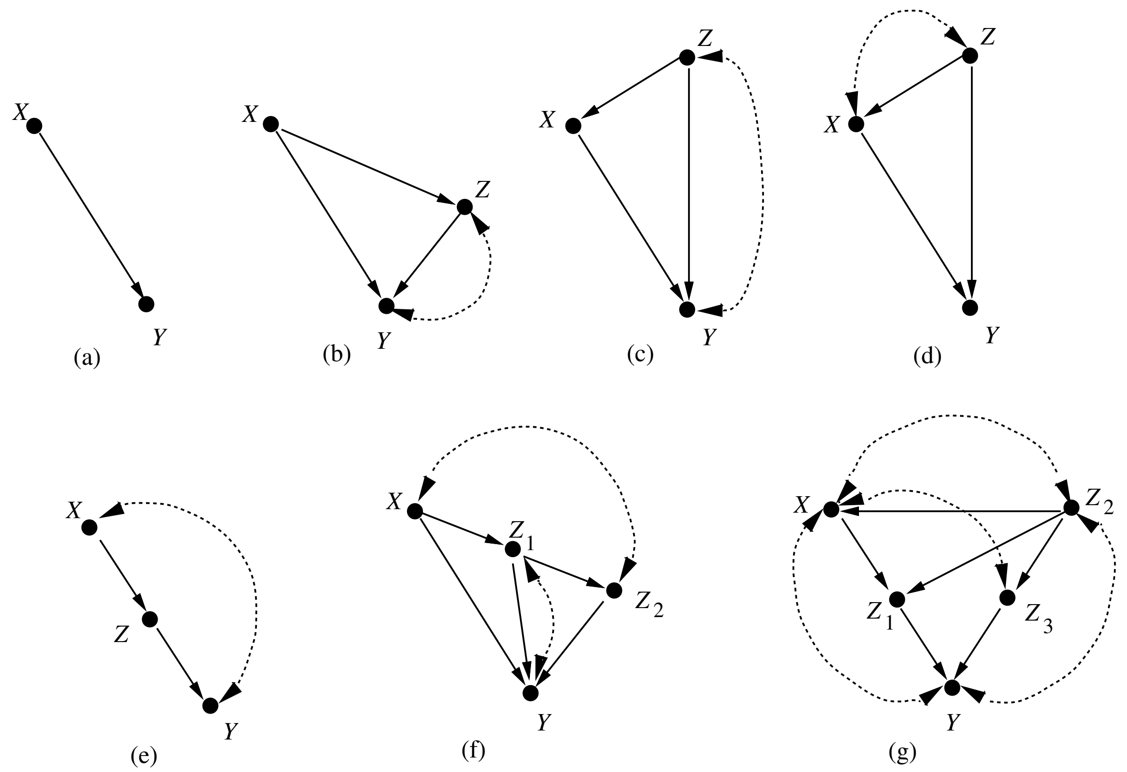

Figure 7:
Typical models in which the total effect of X on Y is identifiable. Dashed lines represent confounding paths, and Z represents observed covariates.

2. 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, $P(y|\hat{x})$ will still be identified from any graph obtained by adding mediating nodes to the diagrams shown in Figure 7.

3. The diagrams in Figure 7 are maximal, in the sense that the introduction of any additional arc or arrow onto an existing pair of nodes would render $P(y|\hat{x})$ no longer identifiable.

4. Although most of the diagrams in Figure 7 contain bow-patterns, none of these patterns emanates from X (as is the case in Figure 8 (a) and (b) below). In general, a necessary condition for the identifiability of $P(y|\hat{x})$ is the absence of a confounding path between X and any of its children on any directed path from X to Y.

5. Diagrams (a) and (b) in Figure 7 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, $P(y|\hat{x}) = P(y|x)$. Likewise, diagrams (c) and (d) in Figure 7 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, $P(y|\hat{x})$ is obtained by a standard adjustment

for Z (as in Eq. (28)):

$$P(y|\hat{x}) = \sum_z P(y|x,z)P(z) \tag{42}$$

6. For each of the diagrams in Figure 7, we can readily obtain a formula for $P(y|\hat{x})$, by using symbolic derivations patterned after those in Section 2.3. The derivation is often guided by the graph topology. For example, diagram (f) in Figure 7 dictates the following derivation. Writing

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

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, $P(z_1, z_2|\hat{x})$ can be obtained from Eq. (38). Likewise, the term $P(y|z_1, z_2, \hat{x})$ can be reduced to $P(y|z_1, z_2, x)$ by Rule 2, since $(Y \perp\!\!\!\perp X|Z_1, Z_2)_{G_{\underline{X}}}$. Thus, we have

$$P(y|\hat{x}) = \sum_{z_1,z_2} P(y|z_1,z_2,x)\ P(z_1|x)\ \sum_{x'} P(z_2|z_1,x')\ P(x') \tag{43}$$

Applying a similar derivation to diagram (g) of Figure 7 yields

$$P(y|\hat{x}) = \sum_{z_1}\sum_{z_2}\sum_{x'} P(y|z_1,\ z_2,\ x')P(x')P(z_1|z_2,\ x)P(z_2) \tag{44}$$

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.

7. In diagrams (e), (f), and (g) of Figure 7, the identifiability of $P(y|\hat{x})$ is rendered feasible through observed covariates, Z, that are affected by the treatment X (i.e., Z being 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]. It is commonly believed [Pratt & Schlaifer 1988] that if a concomitant Z is affected by the treatment, then it should be included in the analysis *only* if we want to learn the conditional effect *given* Z and must be excluded if we want to learn the unconditional total effects. The reason given for the exclusion is that the calculation of total effects often 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 unconditional 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 Eq. (42) (see Eqs. (37), (43), and (44)).

8. Diagrams (b), (c), and (f) of Figure 7 deserve special attention. In each of these graphs, Y has a parent whose effect on Y is not identifiable yet the effect of X on Y is identifiable. This demonstrates that, contrary to linear analysis, 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 of the paths from X to Y.

3.2 Nonidentifying Models

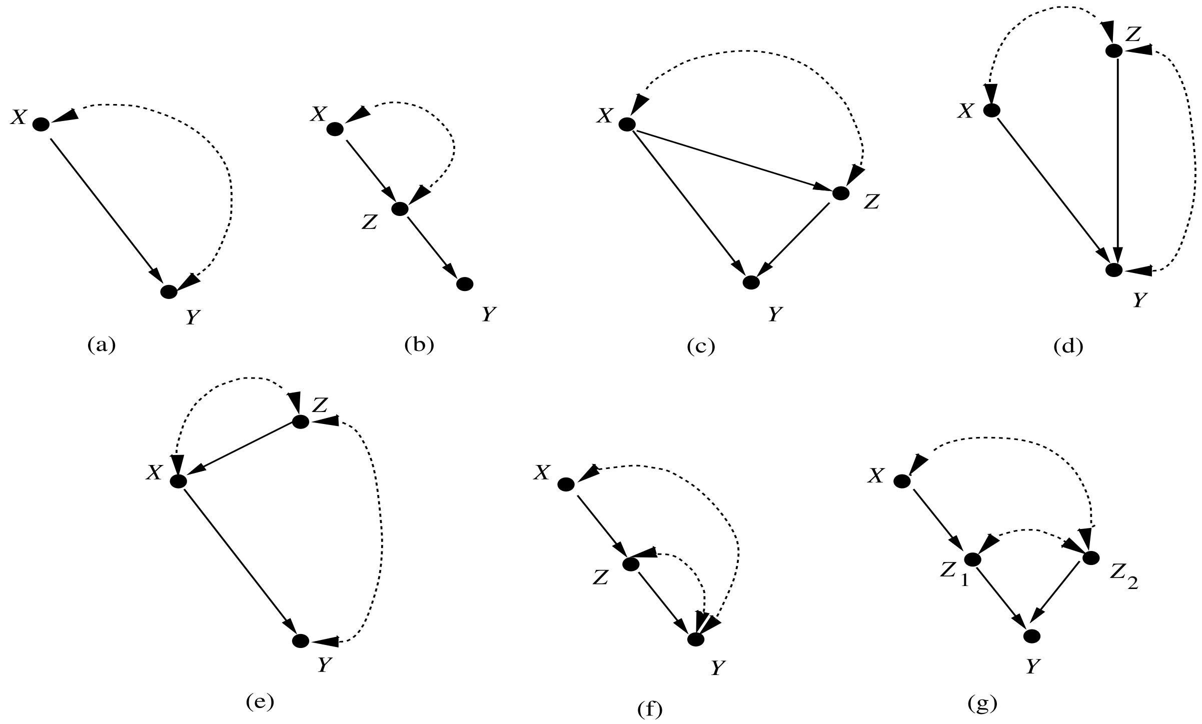

Figure 8:
Typical models in which $P(y|\hat{x})$ is not identifiable.

Figure 8 presents typical graphs in which the total effect of X on Y, $P(y|\hat{x})$, is not identifiable. Noteworthy features of these graphs are as follows.

1. All graphs in Figure 8 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 (see Theorem 3). It is not a sufficient test, though, as is demonstrated by Figure 7 (e), in which the back-door path (dashed) is unblockable and yet $P(y|\hat{x})$ is identifiable.

2. A sufficient condition for the nonidentifiability of $P(y|\hat{x})$ 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 8 (b) and (c). A stronger sufficient condition is that the graph contain any of the patterns shown in Figure 8 as an edge-subgraph.

3. With the exception of (c), all the graphs in Figure 8 are minimal, that is, $P(y|\hat{x})$ is rendered identifiable by removing any arc or arrow from any of these graphs.

4. Graph (g) in Figure 8 does not have any bow-patterns and, moreover, every other causal effect is identifiable except that of X on Y. For example, we can identify $P(z_1|\hat{x})$, $P(z_2|\hat{x})$, $P(y,|\hat{z}_1)$, and $P(y|\hat{z}_2)$, but not $P(y|\hat{x})$. Thus, local identifiability is not sufficient for global identifiability. 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 direct causal effect of one variable on its immediate successor (see Section 3.4).

3.3 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. In such situations, we naturally search for observed covariates that, if adjusted for, would eliminate the confounding effect between X and Y. Such covariates may not always be available, and the question arises whether $P(y|\hat{x})$ can be identified by randomizing a *surrogate* variable Z, which is easier to control than X. More generally, we are interested in a criterion by which a set Z of variables in the diagram can be identified and brought to the investigator's attention as potential surrogates for X.[12] Formally, this problem amounts to transforming $P(y|\hat{x})$ into expressions in which only members of Z obtain the hat symbol.

Diagram (e) in Figure 8 illustrate a simple structure which admits a surrogate experiment. The observed covariate Z is confounded with both X and Y, hence adjusting for Z does not permit the identification of $P(y|\hat{x})$ (i.e., X is not strongly ignorable conditional on Z, by the back-door criterion of Eq. (28)). However, if Z can be controlled by randomized trial, then we can measure $P(x,\ y|\hat{z})$, from which we can compute $P(y|\hat{x})$ using

$$P(y|\hat{x}) = P(y|x,\ \hat{z}) = P(y,\ x|\hat{z})/P(x|\hat{z}) \tag{45}$$

The validity of Eq. (45) can be established by first applying Rule 3 to add $\hat{z}$,

$$P(y|\hat{x}) = P(y|\hat{x},\ \hat{z}) \text{ because } (Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}\overline{Z}}}$$

then applying Rule 2 to exchange $\hat{x}$ with x:

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

The auxiliary diagrams permitting these steps are given in Figure 9.

The use of surrogate experiments is not uncommon. 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.

The derivation leading to Eq. (45) explicates a simple sufficient condition for qualifying a proposed variable Z as a surrogate for X: there must be no direct link from Z to Y and no confounding path between X and Y. 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.

[12]The main distinction between surrogate variables and *instrumental variables* as used in economics [Bowden & Turkington 1984], is that instrumental variables act as though they were randomized while surrogate variables are candidates for randomization. Additionally, the criterion for choosing surrogate variables need not be limited to the standard setting of instrumental variables depicted in Figure 5 (b); it includes any set of variables that would permit (if randomized) the identification of $P(y|\hat{x})$.

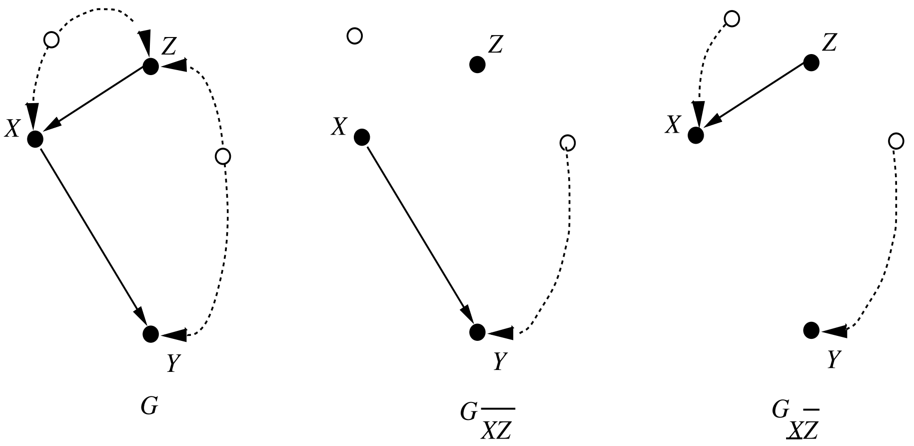

Figure 9:
A causal structure permitting the the identification of $P(y|\hat{x})$ by controlling Z, instead of X.

Note that, according to Eq. (45), only one level of Z suffices for the identification of $P(y|\hat{x})$, for any values of y and x. In other words, Z need not be varied at all, just held constant by external force, and, if the assumptions embodied in G are valid, the r.h.s. of Eq. (45) should attain the same value regardless of the level at which Z is being held constant. In practice, however, several levels of Z will be needed to ensure that enough samples are obtained for each desired value of X. For example, if we are interested in the difference $E(Y|\hat{x}_1) - E(Y|\hat{x}_2)$, then we should choose two values z_1 and z_2 of Z which maximize the number of samples in x_1 and x_2, respectively, and write

$$E(Y|\hat{x}_1) - E(Y|\hat{x}_2) = E(Y|x_1, \hat{z}_1) - E(Y|x_2, \hat{z}_2)$$

Not surprisingly, this expression is equal to the instrumental-variable formula [Angrist et al. 1996]

$$E(Y|\hat{x_1}) - E(Y|\hat{x_2}) = \frac{E(Y|z_1) - E(Y|z_2)}{E(Y|x_1) - E(Y|x_2)}$$

when Z is randomized.

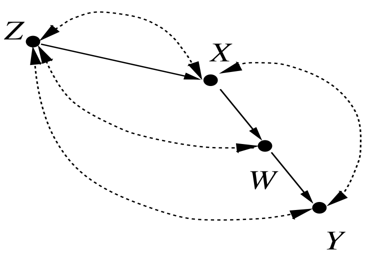

Figure 10:
A more elaborate surrogate experiment; $P(y|\hat{x})$ is identified by controlling Z and measuring W.

Figure 10 illustrates a more general condition for admitting a surrogate experiment. Unlike the condition leading to Eq. (43), randomizing Z now leaves a confounding arc between X and Y. This arc can be neutralized through the mediating variable W, as in the derivation of Eq. (36), and yields the formula

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

Thus, the more general conditions for admitting a surrogate variable Z are:

1. X intercepts all directed paths from Z to Y, and,

2. $P(y|\hat{x})$ is identifiable in $G_{\overline{Z}}$.

3.4 Total, Direct, and Indirect Effects

Path analysis is noted for allowing researchers to decompose the influence of one variable on another into direct, indirect, and total effects [Bollen 1989, page 376]. Yet the path-analytic literature has not been successful in communicating these notions unambiguously to the rest of the scientific community. The standard definition of a total effect is expressed algebraically, in terms of a matrix B of "structural coefficients" [Bollen 1989], and these coefficients, circularly, are defined in terms of total effects when intervening variables are "held constant", [Alwin & Hauser 1975]. With the exception of [Sobel 1990], the notions of "intervening variables", "holding constant", and "structural coefficients" have not been given formal, operational definitions and have remained open to a variety of misinterpretations. Wermuth [1993], for example, interprets "holding X fixed" as "conditioning on X", and finds contradictions in the standard definition of structural equations. Freedman [1987] finds the notion of "fixing" an endogenous variable X to be "self-contradictory", as it conflicts with the assumption that the value of X is functionally determined by the explanatory variables in the equation for X. [Freedman 1987] summarizes the confusion in this area:

> a path model represents the analysis of observational data as if it were the result of an experiment. At points such as this, it would be helpful to know more about the structure of such hypothetical experiments: What is to be held constant, and what manipulated?

To explicate the structure of such hypothetical experiments we need a language in which the notion of "holding constant" is given both formal notation and operational interpretation. The mechanism-based interpretation of "holding constant" as an operation that deletes equations from the model (Definition 3), coupled with the $set(x)$ (or $\hat{x}$) notation introduced in Section 2, constitutes such a language and can be used to provide simple, unambiguous definitions of effect decomposition, for both parametric and nonparametric models.

We start with the general notion of causal effect $P(y|\hat{x})$, as in Definition 3, which applies to arbitrary sets of variables, X and Y. For singleton variables of interest, the notion of causal effect can be specialized to define total and direct effects, as follows.

Definition 5 (*total effect*) *The total effect of X on Y is given by $P(y|\hat{x})$, namely, the distribution of Y while X is held constant at x and all other variables are permitted to run their natural course.*

Definition 6 (*direct effect*) *The direct effect of X on Y is given by $P(y|\hat{x},\ \hat{s}_{XY})$ where S_{XY} is the set of all observed variables in the system, excluding X and Y.*

This definition ascribes to the direct effect the properties of an ideal laboratory; the scientist controls for all possible conditions S_{XY}. It is easy to show (e.g., by applying Rule 3) that there is no need to actually hold *all* other variables constant, since holding constant the direct parents of Y (excluding X) would have the same effect on Y. Thus, we obtain an equivalent definition for direct effect:

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

Readers versed in linear analysis might find it a bit strange that the direct effect of X on Y involves other variables beside X and Y. However, considering that we are dealing with nonlinear interactions, the effect of X on Y should indeed depend on the levels at which we hold the other variables (in the equation for Y). Note also that causal effects are not defined in terms of differences between two expectations, or the relative change in Y with a unit change in X. Such differences can always be determined from the probability distribution $P(y|\hat{x})$. In linear models, for example, the ratio

$$\frac{E(Y|\hat{x}, \hat{pa}_{Y\setminus X}) - E(Y|\hat{x}', \hat{pa}_{Y\setminus X})}{x - x'} = \frac{\partial}{\partial x} E(Y|\hat{x}) = \text{const.}$$

reduces to the ordinary path coefficient between X and Y, regardless of the value taken by $pa_{Y\setminus X}$. In general, if X does not appear in the equation for Y, then $P(y|\hat{x}, \hat{pa}_{Y\setminus X})$ defines a constant distribution on Y, independent of x, which matches our understanding of "having no direct effect". Note also that if PA_Y are not confounded with Y, we can remove the "hat" from the expressions above and define direct effects in terms of ordinary conditional probabilities $P(y|x, pa_{Y\setminus X})$.

The definitions above explicate the operational meaning of structural equations and path coefficients, and should end, I hope, an era of controversy and confusion regarding these entities. Specifically, if G is the graph associated with a set of structural equations, then the assumptions embodied in the equations can be read off G as follows: Every missing arrow, say between X and Y, represents the assumption that X has no causal effect on Y once we intervene and hold the parents of Y fixed. Every missing bi-directed link between X and Y represents the assumption that there are no common causes for X and Y, except those shown in G. Thus, the operational reading of the structural equation $Y = \beta X + \epsilon$ is: "In an ideal experiment where we control X to x and any other set Z of variables (not containing X or Y) to z, Y is independent of z and is given by $\beta x + \epsilon$." The meaning of β 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 and X are correlated (e.g., via another equation $X = \alpha Y + \delta$.) Moreover, this interpretation provides an operational definition for the mystical error-term, ϵ, which is clearly a causal, rather than a statistical, entity.

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

3.5 Evaluating Conditional Policies

The interventions considered thus far were unconditional actions that merely force a variable or a group of variables X to take on some specified value x. In general, interventions may involve 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 $P^*(x|z)$. We will show that computing the effect of such policies is equivalent to computing the expression $P(y|\hat{x}, z)$.

Let $P(y|set(X = g(Z)))$ stand for the distribution (of Y) prevailing under the policy $(X = g(Z))$. To compute $P(y|set(X = g(Z)))$, we condition on Z and write

$$
\begin{aligned}
&P(y|set(X = g(Z))) \\
&\quad = \sum_z P(y|set(X = g(z)), z) P(z|set(X = g(z))) \\
&\quad = \sum_z P(y|\hat{x}, z)|_{x=g(z)} P(z) \\
&\quad = E_z[P(y|\hat{x}, z)|_{x=g(z)}]
\end{aligned}
$$

where the equality

$$P(z|set(X = g(z))) = P(z)$$

stems from the fact that Z cannot be a descendant of X, hence, whatever control one exerts on X, it can have no effect on the distribution of Z. Thus, we see that the causal effect of a policy $X = g(Z)$ can be evaluated directly from the expression of $P(y|\hat{x}, z)$, simply by substituting $g(z)$ for x and taking the expectation over Z (using the observed distribution $P(z)$).

The identifiability condition for policy intervention is somewhat stricter than that for a simple intervention. Clearly, whenever a policy $set(X = g(Z))$ is identifiable, the simple intervention $set(X = x)$ is identifiable as well, as we can always get the latter by setting $g(Z) = X$. The converse, does not hold, however, because conditioning on Z might create dependencies that will prevent the successful reduction of $P(y|\hat{x}, z)$ to a hat-free expression.

A stochastic policy, which imposes a new conditional distribution $P^*(x|z)$ for x, can be handled in a similar manner. We regard the stochastic intervention as a random process in which the unconditional intervention $set(X = x)$ is enforced with probability $P^*(x|z)$. Thus, given $Z = z$, the intervention $set(X = x)$ will occur with probability $P^*(x|z)$ and will produce a causal effect given by $P(y|\hat{x}, z)$. Averaging over x and z gives

$$P(y|P^*(x|z)) = \sum_x \sum_z P(y|\hat{x}, z)P^*(x|z)P(z)$$

Since $P^*(x|z)$ is specified externally, we see again that the identifiability of $P(y|\hat{x}, z)$ is a necessary and sufficient condition for the identifiability of any stochastic policy that shapes the distribution of X by the outcome of Z.

It should be noted, however, that in planning applications the effect of an action may be to invalidate its preconditions. To represent such actions, temporally indexed causal networks may be necessary [Dean & Kanawaza 1989] or, if equilibrium conditions are required, cyclic graphs can be used [Balke & Pearl 1995].

4 Discussion

This paper demonstrates that:

1. The effect of intervening policies often be identified (from nonexperimental data) without resorting to parametric models.

2. The conditions under which such nonparametric identification is possible can be determined by simple graphical criteria.

3. When the effect of interventions is not identifiable, the causal graph may suggest non-trivial experiments which, if performed, would render the effect identifiable.

While the ability to assess the effect of interventions from nonexperimental data has many applications in the social and health sciences, perhaps the most practical result reported in this paper is the solution of the long standing problem of covariate adjustment. The reader might recognize Eq. (42) as the standard formula for covariate adjustment (also called "stratification"), which is used both for improving precision and for minimizing confounding bias. However, a formal, general criterion for deciding whether a set of covariates Z qualifies for adjustment has long been wanting. In the context of linear regression models, the problem amounts to deciding whether it is appropriate to add a set Z of variables to the regression of Y on X. Most of the statistical literature is satisfied with informal warnings that "Z should be quite unaffected by X" [Cox 1958, page 48], which is necessary but not sufficient (see Figure 8(d)) or that X should not precede Z [Shafer 1996, page 326], which is neither necessary nor sufficient. In some academic circles, a criterion called "ignorability" is invoked [Rosenbaum & Rubin 1983], which merely paraphrases the problem in the language of counterfactuals. Simplified, ignorability reads: Z is an admissible covariate relative to the effect of X on Y if, for every x, the value that Y would obtain had X been x is conditionally independent of X, given Z (see appendix II for further discussion of counterfactual analysis). In contrast, Eq. (26) provides an admissibility test which is both precise and meaningful, as it is applicable directly to the elementary processes (i.e., linkages in the graph) around which scientific knowledge is organized. This test (called the "back-door criterion" in [Pearl 1993]) reads: Z is an admissible set of covariates relative to the effect of X on Y if:

(i) no node in Z is a descendant of X, and

(ii) Z d-separates X from Y along any path containing an arrow into X (equivalently, $(Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}}}$).

We see, for instance, that Z qualifies as admissible covariates relative the effect of X on Y in Figure 7(d) but not in Figure 8(d). The graphical definition of admissible covariates replaces statistical folklore with formal procedures, and should enable analysts to systematically select an optimal set of observations, namely, a set Z that minimizes measurement cost or sampling variability.

It is important to note several limitations and extensions of the method proposed in this paper. First, the structural models discussed so far consist only of behavioral equations; definitional equalities and equilibrium constraints are excluded, as these do not respond to intervention in the manner described in Definition 3. One way of handling mixtures of behavioral and equilibrium equations is to treat the latter as observational events, on which to condition the probabilities [Strotz & Wold 1960]. For example, the econometric equilibrium constraint $q_d = q_s$, equating quantity demanded and quantity supplied, would be treated by adding a "dummy" behavioral equation

$$S = q_s - q_d$$

(S connoting "stock growth") and, then, conditioning the resulting probabilities on the event $S = 0$. Such conditioning events tend to introduce new dependencies among the variables in the graphs, as dictated by the d-separation criterion. Consequently, in applying the inferences rule of Theorem 3, one would have to consult graphs in which the dummy variables have been permanently conditioned.

A second extension concerns the use of the causal calculus (Theorem 1) in cyclic models. The subtheory interpretation of control queries (Definition 3) carries over to cyclic systems [Strotz & Wold 1960, Sobel 1990], but then two issues must be addressed. First, the analysis of identification is meaningful only when the resulting system is stable. Therefore, we must modify the definition of identifiability by considering only the set of stable theories for each structural model and for each submodel [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 non-recursive linear models and extended, using an augmented graph to any arbitrary set of stable equations [Spirtes 1994]. However, the computation of control queries will be harder in cyclic networks, because complete reduction of control queries to hat-free expressions may require the solution of nonlinear equations.

Having obtained nonparametric formulas for causal effects does not imply, of course, that one should refrain from using parametric forms in the estimation phase of the study. When data are scarce, prior information about shapes of distributions and the nature of causal interactions can be extremely useful, and it can be incorporated into the analysis by limiting the distributions in the estimand formulas to whatever parametric family of functions are deemed plausible by the investigator. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable in a given problem, then nonparametric formulas of the type (see Eq. (33))

$$P(y|\hat{x}) = \sum_z P(y|x,\ z)P(z) \tag{46}$$

will be converted to

$$E(Y|\hat{x}) = \int_y \int_z \ yf(y|x,\ z)f(z)dydz = R_{yx\cdot z}x \tag{47}$$

and the estimation problem reduces to that of estimating (e.g., by least-squares) the regression of Y on X and Z. Similarly, the estimand given in Eq. (37) can be converted to a product

$$E(Y|\hat{x}) = R_{xz}\beta_{zy\cdot x}x \tag{48}$$

where $\beta_{zy\cdot x}$ is the standardized regression coefficient. More sophisticated estimation techniques, tailored specifically for causal inference can be found in [Robins 1992].

Finally, a few comments regarding the notation introduced in this paper. 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 (see Appendix II).

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 can be anticipated, therefore, that by separating issues of identification and parametric form this article should serve to make the language of path analysis more accessible to the scientific community (see discussions following [Pearl 1995a]).

Acknowledgment

This investigation benefitted from discussions with Joshua Angrist, Peter Bentler, David Cox, Sander Greenland, Arthur Goldberger, David Hendry, Paul Holland, Guido Imbens, Ed Leamer, Rod McDonald, John Pratt, James Robins, Paul Rosenbaum, Donald Rubin, and Michael Sobel. The research was partially supported by Air Force grant #AFOSR 90 0136, NSF grant #IRI-9200918, and Northrop-Rockwell Micro grant #93-124.

APPENDIX I. Smoking and the Genotype Theory: An Illustration

To illustrate the usage of the causal effects computed in Subsection 2.3, we will associate the model of Figure 1 with a concrete example concerning the evaluation of the effect of smoking (X) on lung cancer (Y). According to many, the tobacco industry has managed to stay anti-smoking legislation by arguing that the observed correlation between smoking and lung cancer could be explained by some sort of carcinogenic genotype (U) which involves inborn craving for nicotine.[13]

The amount of tar (Z) deposited in a person's lungs is a variable that promises to meet the conditions specified by the structure of Figure 1. To justify the missing link between X and Y, we must assume that smoking cigarettes (X) has no effect on the production of lung cancer (Y) except that mediated through tar deposits. To justify the missing link between U and Z, we must assume that, even if a genotype is aggravating the production of lung cancer, it nevertheless has no effect on the amount of tar in the lungs except indirectly, through cigarette smoking.

To demonstrate how we can assess the degree to which cigarette smoking increases (or decreases) lung cancer risk, we will construct a hypothetical study in which the three variables, $X, Y,$ and Z, were measured simultaneously on a large, randomly selected sample from the population. To simplify the exposition, we will further assume that all three variables are binary, taking on true (1) or false (0) values. A hypothetical data set from a study on the relations among tar, cancer, and cigarette smoking is presented in Table 1.

	Group Type	$P(x, z)$ Group Size (% of Population)	$P(Y = 1 \mid x, z)$ % of Cancer Cases in Group
$X = 0,\ Z = 0$	Non-smokers, No tar	47.5	10
$X = 1,\ Z = 0$	Smokers, No tar	2.5	90
$X = 0,\ Z = 1$	Non-smokers, Tar	2.5	5
$X = 1,\ Z = 1$	Smokers, Tar	47.5	85

Table 1

The table shows that 95% of smokers and 5% of non-smokers have developed high levels of tar in their lungs. Moreover, 81.51% of subjects with tar deposits have developed lung cancer, compared to only 14% among those with no tar deposits. Finally, within each of the two groups, tar and no tar, smokers show a much higher percentage of cancer than non-smokers do.

These results seem to prove that smoking is a major contributor to lung cancer. However, the tobacco industry might argue that the table tells a different story—that smoking actually decreases, not increases, one's risk of lung cancer. Their argument goes as follows. If you decide to smoke, then your chances of building up tar deposits are 95%, compared to 5% if you decide not to smoke. To evaluate the effect of tar deposits, we look separately at two groups, smokers and non-smokers. The table shows that tar deposits have a protective effect in both groups: in smokers, tar deposits lower cancer rates from 90% to 85%; in non-smokers, they lower cancer rates from 10% to 5%. Thus, regardless of

[13] For an excellent historical account of this debate, see [Spirtes et al. 1993, pp. 291–302].

whether I have a natural craving for nicotine, I should be seeking the protective effect of tar deposits in my lungs, and smoking offers a very effective means of acquiring them.

To settle the dispute between the two interpretations, we note that, while both arguments are based on stratification, the anti-smoking argument invokes an illegal stratification over a variable (Z) that is affected by the treatment (X). The tobacco industry's argument, on the the hand, is made up of two steps, neither of which involves stratification over treatment-affected variables: stratify over smoking to find the effect of tar deposit on lung cancer, then average (not stratify) over tar deposits when we consider each of the decision alternatives, smoking vs. non-smoking. This is indeed the intuition behind the formula in Eq. (32) and, given the causal assumptions of Figure 7, the tobacco industry's argument is the correct one (see [Pearl 1995a, Pearl 1994] for formal derivation).

To illustrate the use of Eq. (32), let us use the data in Table 1 to calculate the probability that a randomly selected person will develop cancer ($y_1 : Y = 1$) under each of the following two actions: smoking ($x_1 : X = 1$) or not smoking ($x_0 : X = 0$).

Substituting the appropriate values of $P(y|x)$, $P(y|x, z)$, and $P(x)$ gives

$$\begin{aligned} E[P(y_1|x_1, u)] &= .05(.10 \times .50 + .90 \times .50) + .95(.05 \times .50 + .85 \times .50) \\ &= .05 \times .50 + .95 \times .45 = .4525 \\ E[P(y_1|x_0, u)]) &= .95(.10 \times .50 + .90 \times .50) + .05(.05 \times .50 + .85 \times .50) \\ &= .95 \times .50 + .05 \times .45 = .4975 \end{aligned} \tag{49}$$

Thus, contrary to expectation, the data prove smoking to be somewhat beneficial to one's health.

The data in Table 1 are obviously unrealistic and were deliberately crafted so as to support the genotype theory. However, this exercise was meant to demonstrate how reasonable qualitative assumptions about the workings of mechanisms can produce precise quantitative assessments of causal effects when coupled with nonexperimental data. In reality, we would expect observational studies involving mediating variables to refute the genotype theory by showing, for example, that the mediating consequences of smoking, such as tar deposits, tend to increase, not decrease, the risk of cancer in smokers and non-smokers alike. The estimand given in Eq. (32) could then be used for quantifying the causal effect of smoking on cancer.

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 [Rubin 1990] 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 equation models. Consider a set T of equations

$$X_i = f_i(PA_i, U_i) \quad i = 1, \ldots, n \tag{50}$$

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). (50) is similar to (14), 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 subtheory 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

$$Y(x,u) \triangleq Y_{T_x}(u) \tag{51}$$

namely, $Y(x,u)$ is the (unique) solution of Y under the realization $U = u$ in the subtheory 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. (51) 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 vector of 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

$$X = x \implies Y(x) = Y \tag{52}$$

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 5(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 4 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 (50). 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

$$Y(pa_Y) = Y(pa_Y, s) \tag{53}$$

2. Independence restrictions: If $Z_1, \ldots, Z_k$ is any set of nodes not connected to Y via dashed arcs, we have

$$Y(pa_Y) \perp\!\!\!\perp \{Z_1(pa_{Z_1}), \ldots, Z_k(pa_{Z_k})\} \tag{54}$$

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. (52)) 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:

$$\textbf{Degeneracy}: \quad Y(\emptyset) = Y \tag{55}$$

$$\textbf{Composition}: \quad Y(x) = Y(x, Z(x)) \text{ for any } Z \text{ disjoint of } \{X, Y\} \tag{56}$$

$$\textbf{Sure}-\textbf{thing}: \quad \text{If } Y(x, z) = Y(x', z)\ \forall\ x' \neq x, \text{ then } Y(x, z) = Y(z) \tag{57}$$

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[14] 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 antecedent $X = x$ is redundant, namely, we need not concern ourselves with the value that X actually obtains.

Properties (56)-(57) are theorems in the structural interpretation of $Y(x,u)$ as given in Eq. (51) [Galles & Pearl 1997]. 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". It is easy to verify that composition and degeneracy imply the consistency rule of (52); substituting $X = \{\emptyset\}$ in (59) yields $Y = Y(z)$ if $Z = z$, which is equivalent to (52).

As an example, let us compute the causal effects associated with the model shown in Figure 2 (or Eqs. (1)-(3)). The parents sets a given by:

$$PA_X = \{\emptyset\},\ PA_Z = \{X\},\ PA_Y = \{Z\} \tag{58}$$

Consequently, the exclusion restrictions (53) translate into:

$$\begin{aligned} Z(x) &= Z(y,x) & (59) \\ X(y) &= X(z,y) = X(z) = X & (60) \\ Y(z) &= Y(z,x) & (61) \end{aligned}$$

The independence restrictions (54) translate into:

$$Z(x) \perp\!\!\!\perp \{Y(z), X\} \tag{62}$$

Task-1, compute $P^*(Z(x) = z)$ (Equivalently $P(z|\hat{x})$)
From (62) we have $Z(x) \perp\!\!\!\perp X$, hence

$$P^*(Z(x) = z) = P^*(Z(x) = z|x) = P^*(z|x) = P(z|x) \tag{63}$$

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

$$P^*(Y(z) = y) = \sum_x P^*(Y(z) = y|x)P^*(x) \tag{64}$$

[14] This axiom was communicated to me by James Robins (1995, in conversation) as a property needed for defining a structure he calls "finest fully randomized causal graphs" [Robins 1986, pp. 1419–1423]. In Robins' analysis, $Y(x,z)$ and $Z(x)$ may not be defined.

From (62) we have

$$Y(z) \perp\!\!\!\perp Z(x)|X \tag{65}$$

hence

$$\begin{aligned} P^*(Y(z)=y|x) &= P^*(Y(z)=y|x, Z(x)=z) && \text{by (52)} \\ &= P^*(Y(z)=y|x,z) && \text{by (40)} \\ &= P^*(y|x,z) && \text{by (40)} \\ &= P(y|x,z) \end{aligned} \tag{66}$$

Substituting (66) in (64), gives

$$P^*(Y(z)=y) = \sum_x P(y|x,z)P(x) \tag{67}$$

which is the celebrated covariate-adjustment formula for causal effect, as in Eq. (42).

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 (61)), 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 (67) and (63), 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 (51) and, conversely, counterfactual analysts can use the constraints (over P^*) given by Eqs. (53) and (54) as a definition of graphs, structural equations and the physical processes which they represent.

References

[Alwin & Hauser 1975] Alwin, D.F., and Hauser, R.M., "The decomposition of effects in path analysis," *American Sociological Review*, 40, 37–47, 1975.

[Angrist et al. 1996] Angrist, J.D., Imbens, G.W., and Rubin, D.B., "Identification of causal effects using instrumental variables (with Comments)," *Journal of the American Statistical Association*, 91(434), 444–472, June 1996.

[Balke & Pearl 1993] Balke, A., and Pearl, J., "Nonparametric bounds on causal effects from partial compliance data," Department of Computer Science, University of California, Los Angeles, Technical Report R-199, 1993. To appear in *Journal of the American Statistical Association.*

[Balke & Pearl 1994] Balke, A., and Pearl, J., "Counterfactual probabilities: Computational methods, bounds, and applications," In R. Lopez de Mantaras and D. Poole (Eds.), *Uncertainty in Artificial Intelligence - 10*, Morgan Kaufmann, San Mateo, 46–54, 1994.

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

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

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

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

[Dean & Kanawaza 1989] Dean, T., and Kanazawa, K., "A model for reasoning about persistence and causation," *Computational Intelligence*, 5, 142–150, 1989.

[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, with Tinbergen's comments, in *Autonomy of Economic Relations*, Oslo: Universitetets Socialokonomiske Institutt, 1948). Cited in M.S. Morgan, *The History of Econometric Ideas*, Cambridge University Press, Cambridge, 1990.

[Galles & Pearl 1995] Galles, D., and Pearl, J., "Testing identifiability of causal effects," In P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann, San Francisco, CA, 185-195, 1995.

[Galles & Pearl 1997] Galles, D., and Pearl, J., "Axioms of causal relevance," Preliminary version in *Proceedings of the Fourth International Conference on Mathematics and AI*, Fort Lauderdale, FL, 64–67, January, 1996. Revised May 1997. To appear in *Artificial Intelligence*.

[Geiger 1990] Geiger, D., "Graphoids: A qualitative framework for probabilistic inference," UCLA, Ph.D. Dissertation, Computer Science Department, 1990.

[Geiger & Pearl 1988] Geiger, D., and Pearl, J., "On the logic of causal models," *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*, St. Paul, MN, 136–147, 1988. Also in L. Kanal et al. (Eds.), *Uncertainty in Artificial Intelligence*, 4, Elsevier Science Publishers, North-Holland, Amsterdam, 3–14, 1990.

[Geiger et al. 1990] Geiger, D., Verma, T.S., and Pearl, J., "Identifying independence in Bayesian networks." *Networks*, 20, 507–534, 1990.

[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.

[Imbens & Angrist 1994] Imbens, G.W. and Angrist, J.D., "Identification and estimation of local average treatment effects," *Econometrica*, 62(2), 467–475, 1994.

[Koopman & Reiersol, 1950] Koopman, T.C. and Reiersol, O., "The identification of structural characteristics," *Annals of Mathematical Statistics*, 21, 165–181, 1950.

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

[Marschak 1953] Marschak, J., "Economic measurements for policy and prediction," in T. Koopmans and W. Hood (Eds.), *Studies in Econometric Method, Cowles Commission Monograph 14*, Chapter 1, Yale University Press, New Haven, 1953.

[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, 1988.

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

[Pearl 1994] Pearl, J., "A probabilistic calculus of actions," in R. Lopez de Mantaras and D. Poole (Eds.), *Uncertainty in Artificial Intelligence* 10, Morgan Kaufmann, San Mateo, pp. 454-462, 1994.

[Pearl 1995a] Pearl, J., "Causal diagrams for experimental research, (with discussion)" *Biometrika*, 82(4) 669–710, 1995.

[Pearl 1995b] Pearl, J., "Causal inference from indirect experiments," *Artificial Intelligence in Medicine Journal*, 7, 561–582, 1995.

[Pearl 1995c] Pearl, J., "On the testability of causal models with latent and instrumental variables," In P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann Publishers, San Francisco, 435–443, 1995.

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

[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: Proceeding of the Second International Conference*, Morgan Kaufmann, San Mateo, 1991, pp. 441–452.

[Pearl et al. 1990] Pearl, J., Geiger, D., and Verma, T., "The logic of influence diagrams, in R.M. Oliver and J.Q. Smith (Eds.), *Influence Diagrams, Belief Nets and Decision Analysis*, John Wiley and Sons, New York, 67–87, 1990.

[Pratt & Schlaifer 1988] Pratt, J., and Schlaifer, R., "On the interpretation and observation of laws," *Journal of Economics*, 39, 23–52, 1988.

[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 Modeling*, Vol. 7, 1393–1512, 1986.

[Robins 1987] Robins, J.M., "A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods," *Journal of Chronic Diseases*, 40, Suppl. 2, 139S–161S, 1987.

[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, 113–159, 1989.

[Robins 1992] Robins, J., "Estimation of the time-dependent accelerated failure time model in the presence of confounding factors," *Biometrika*, 79(2), 321–334, 1992.

[Robins & Greenland 1992] Robins, J.M., and Greenland, S., "Identifiability and exchangeability for direct and indirect effects," *Epidemiology*, 3(2), 143–155, 1992.

[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*, 3(4), 319–336, 1992.

[Rosenbaum 1984] Rosenbaum, P.R., "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(5), 656–666, 1984.

[Rosenbaum 1989] Rosenbaum, P.R., "The role of known effects in observational studies," *Biometrics*, 45, 557–569, 1989.

[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 1990] Rubin, D.B., "Formal modes of statistical inference for causal effects," *Journal of Statistical Planning and Inference*, 25, 279–292, 1990.

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

[Simon 1977] Simon, H.A., *Models of Discovery: and Other Topics in the Methods of Science*, D. Reidel, Dordrecht, Holland, 1977.

[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.

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

[Spiegelhalter et al. 1993] Spiegelhalter, D.J., Lauritzen, S.L., Dawid, P.A., and Cowell, R.G., "Bayesian analysis in expert systems," *Statistical Science*, 8(3), 219–247, 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.

[Verma 1990] Verma, T.S., "Causal networks: Semantics and expressiveness," in R. Shachter et al. (Eds.), *Uncertainty in Artificial Intelligence*, 4, Elsevier Science Publishers, North-Holland, Amsterdam, 69–76, 1990.

[Verma & Pearl 1990] Verma, T.S. and Pearl, J., "Equivalence and synthesis of causal models." In *Uncertainty in Artificial Intelligence 6*, Elsevier Science Publishers, Cambridge, 220-227, 1990.

[Wermuth 1993] Wermuth, N., "On block-recursive regression equations" (with discussion), *Brazilian Journal of Probability and Statistics*, 6, 1–56, 1992.