UCLA

Department of Statistics Papers

Title

On the Identification of Nonparametric Structural Models

Permalink

<https://escholarship.org/uc/item/4bq5r0c4>

Author

Judea Pearl

Publication Date

2011-10-25

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 identifiability 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 parameters but rather uniqueness of the set of predictions of interest to the investigator. For example, predicting the effects of changes, interventions, and control.

We provide sufficient and necessary conditions for identifying a set of causal 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 verified qualitatively, by inspection, using the graphical representation of the structural model. When compared to standard identifiability tests of linear models (e.g., rank and order), our criteria serve to segregate tests that hold universally from those that are artifacts of the linearity assumption.

1 Introduction

Consider the following set of structural equations:

$$X = f_1(U) \tag{1}$$

$$Z = f_2(X, V) \tag{2}$$

$$Y = f_3(Z, U, W) \tag{3}$$

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 modeller is not willing to commit to any particular functional form but feels strongly about the qualitative nature of the interactions—for example, that X need not be an argument of f_3 , that U should be an argument of f_3 , and so on. Graphically, these interactions can be represented by the path diagram of Figure 1. Note that the arcs in Figure 1 will be labeled with the functions themselves, not with coefficients as in traditional path analysis which allows only linear interactions.

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 parameters of the model to the greatest extent possible.

When the equations are parametric, then, given the right set of assumptions on the disturbances, it is often possible to estimate all relevant parameters; such equations are called “identifiable”, that is, they define a unique transformation between the disturbances $\{U, V, W\}$ and the observables $\{X, Z, Y\}$. However, 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, for example, Eq. (2), $Z = f_2(X, V)$, assuming that V is independent of X , having a known distribution $P(v)$. Assuming further that we are given the joint distribution $P(x, z)$ of X and Z , we are still

unable to determine the functional form of f_2 . 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 doomed to failure and that nothing useful can be inferred from such loose specification of structural equations.

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 distribution $P(x, z, y)$ which, if needed, can be represented by the structural form of Eqs. (1)-(3). For example, assuming that X and Z are binary, and considering only Eqs. (1) and (2), we can arbitrarily choose U and V to be uniformly distributed between 0 and 1, write

$$f_1(U) = \begin{cases} 1 & \text{if } U \leq \alpha \\ 0 & \text{if } U > \alpha \end{cases} \quad (4)$$

$$f_2(X, V) = \begin{cases} 1 & \text{if } X = 1, V \leq \beta \\ 1 & \text{if } X = 0, V \leq \gamma \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

and adjust the constants α , β , and γ to match the observed probabilities:

$$\alpha = P(X = 1) \quad \beta = P(Z = 1|X = 1) \quad \gamma = P(Z = 1|X = 0) \quad (6)$$

By this choice of functions and distributions, we guarantee that samples generated by the first two equations will satisfy the joint distribution $P(x, z)$. Although obviously the choice is not unique, it is nevertheless compatible with the observations and will yield the same probabilistic predictions as would the joint distribution $P(x, z)$.

If every joint distribution P can be simulated in this way, by some parametrization of the model, we say that our model is *universal*; if not, we say that it is *falsifiable*, because potentially some empirical evidence capable of proving our model wrong exists. As an example of a falsifiable model, consider Eqs. (1)-(3) with Eq. (3) modified to read $Y = f_3(Z, W)$. In this case, the assumption of mutual independence among U, V , and W dictates that X and Y be conditionally independent given Z . Since not every joint distribution $P(x, z, y)$ satisfies this independence, the model is falsifiable. The model defined by Eqs. (1)-(3), on the other hand, happens to be universal but, as will become clear, the proof is not trivial and we do not know of any simple general test for falsifiability.

Falsifiable models, if they correctly embody the constraints specific to a given domain, can be very effective in providing a parsimonious description of the observed joint distribution function. However, the primary value of structural equations lies not in summarizing distribution functions but in encoding causal information. This, in turn, enables us to

answer questions about control, such as “What would be the expected value of Y if we were to intervene and set the value of X to x ”, which is different, of course, from the observational query “What would be the expected value of Y if we were to find X at level x ”. Observational queries can be answered directly from the joint distribution $P(x, y, z)$, while control queries require a causal model, 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

$$E(Y|\hat{x}) = E[Y|set(X = x)] \quad (7)$$

for the controlled expectation and

$$E(Y|x) = E(Y|X = x) \quad (8)$$

for the standard conditional 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 linearized version of the model defined by Eqs. (1)-(3) could be obtained from the product of path coefficients along the path $X \longrightarrow Z \longrightarrow Y$. In the nonlinear 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 , as it requires the solution (for the expectation of Y) of a modified set of equations in which f_1 is “wiped out” [Strotz & Wold 1960, Sobel 1990] and X is replaced by the constant x :

$$Z = f_2(x, V) \quad (9)$$

$$Y = f_3(Z, U, W) \quad (10)$$

This modified model represents a local intervention $X = x$ that overrules the previous connection between X and U and leaves the other equations unaltered. Similar modifications of the model are required for the computation of $E(Z|\hat{x})$, $E(X|\hat{z})$, or $E(X|\hat{y})$, which can easily be shown to yield $E(Z|\hat{x}) = E(Z|x)$, $E(X|\hat{z}) = E(X)$, and $E(X|\hat{y}) = E(X)$. By contrast, 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 involves more work (see Section 2.3), but nevertheless $E(Y|\hat{x})$ can be expressed in terms of the observed joint probability $P(x, y, z)$.

What then would be an appropriate definition of “identifiability” for nonparametric models? The answer depends upon the use to which the model is being put. If we take the traditional stand that the purpose of structural modeling is to provide answers to control queries [Haavelmo 1943, Marschak 1953], then a reasonable definition of identifiability is that such answers are *unique*. Accordingly, we 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. 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; this model permits the identification of all control parameters. For example,

1. $P(x|\hat{y}, \hat{z}) = P(x)$
because U is exogenous, so no other variable has influence over X ; and
2. $P(z|\hat{x}) = P(z|x)$
because V is independent of X , so Eq. (2) is a regression equation.

$P(y|\hat{x})$ is harder to compute, but it is identifiable and is given by

3. $P(y|\hat{x}) = \sum_z [P(z|x) \sum_{x'} P(y|x', z) P(x')]$.

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 parameters above.

As an example of a nonidentifiable system, consider replacing Eq. (2) with the equation $Z = f_2(X, V, U)$ instead of $f_2(X, V)$. This change allows us to emulate $P(x, z)$ with many additional choices of functions f_1 and f_2 , yielding different values for $P(z|\hat{x})$. For example, consider the following choice of $P(u), f_1$, and f_2 :

$$\begin{aligned}
 U &\in \{1, 2, 3, 4\} \\
 P(u) &= \begin{cases} P(X=0, Z=0) & \text{if } U=0 \\ P(X=0, Z=1) & \text{if } U=1 \\ P(X=1, Z=0) & \text{if } U=2 \\ P(X=1, Z=1) & \text{if } U=3 \end{cases} \quad (11)
 \end{aligned}$$

$$f_1(U) = \begin{cases} 1 & \text{if } U=2 \text{ or } 3 \\ 0 & \text{if } U=0 \text{ or } 1 \end{cases} \quad (12)$$

$$f_2(X, V, U) = \begin{cases} 1 & \text{if } U=1 \text{ or } 3 \\ 0 & \text{if } U=0 \text{ or } 1 \end{cases} \quad (13)$$

This model clearly is faithful to $P(x, z)$, as is the one in Eqs. (4)-(6). However, this choice yields $P(z|\hat{x}) = P(z)$, because f_2 is not a function of X (i.e., the dependence between X and Z is fully attributed to their common cause U), while the model in Eqs. (4)-(6) gives $P(z|\hat{x}) = P(z|x)$, because V is independent of X . Thus, the modified model, with $Z = f_2(X, V, U)$, is not identifiable.

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

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

$$X_i = f_i(X_1, X_2, \dots, X_{i-1}; U_1, \dots, U_m), \quad i = 1, 2, \dots, n, \quad (14)$$

where $X_1, \dots, X_n$ are observed variables, $U_1, \dots, U_m$ are unobserved (or latent) disturbances.¹ 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 equations² 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, modellers 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

¹The recursive nature of Eq. (14) corresponds to a lower triangular matrix B in linear models [Bollen 1989] and therefore excludes reciprocal causation. This restriction is not essential to the main thrust of the discussion and is introduced here to simplify the exposition. Likewise, 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. (14).

²In linear models, these correspond to the “zero-coefficient” restrictions, while the independence restrictions 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. (14) (equivalently, by the corresponding graph $G(M)$) delineates a set of *parameterized models* which we call *theories*. Each theory $T = \langle \{f_i\}, P(\mathbf{u}) \rangle$ in M corresponds to a specific choice of function f_i and a specific choice of disturbance distribution $P(\mathbf{u}) = P(u_1, \dots, u_m)$, both satisfying the restrictions imposed by M . For each theory T of M , there is a corresponding unique probability distribution $P_T(\mathbf{x}) = P_T(x_1, \dots, x_n)$, which we say to be “generated” by T .

Definition 1 *We say that $P(\mathbf{x})$ is compatible with M iff there exists a theory T of M such that*

$$P(\mathbf{x}) = P_T(\mathbf{x})$$

A model M is said to be universal if it is compatible with every arbitrary $P(\mathbf{x})$; otherwise, it is said to be falsifiable. $\square$

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 $P(\mathbf{x})$, using mutually independent disturbances. Section 1 provides additional examples of universal and falsifiable models.

2.1.2 Queries and Identifiability

A query q is a function from structural theories to the real numbers. For example, the queries

$$\begin{aligned} q_1 : & \quad f_2(X = 1, V = 3.06) = ? \\ q_2 : & \quad P(U = 1 | Y = 0.8) = ? \\ q_3 : & \quad P(X = 1 | Y = 3) = ? \\ q_4 : & \quad E_u[P(Y = 1 | X = 1, u)] = ? \end{aligned} \tag{15}$$

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 can be computed uniquely. Note, moreover, that, in view of the structural restriction that $\{U, V, W\}$ be independent, queries q_3 and q_4 (but not q_1 or q_2) will have a unique answer in all theories that generate a given distribution $P(x, y, z)$.³ 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 ,*

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

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

³The reader will recognize q_4 to be equal to $P(Y = 1 | \text{set}(X = 1))$ discussed in Section 1.

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 could be generated by some theory of M .

Technically, the reason for restricting the observed distributions to positive distributions is to avoid conditioning on events with zero probabilities. Conceptually, this restriction means that identifiability is defined under conditions where every f_i is stochastic.

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 $P(\mathbf{x})$. 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 $Q = Q_2$ 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. (9)-(10) 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 function from the theories of M to the reals, defined by

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

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

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 brings 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].⁴ An explicit translation of interventions to “wiping out” equations

⁴The notion of invariance, and its operational derivative of Definition 3 can, in fact, be taken as 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

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 1993c].

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 $\mathbf{X} = \{X_1, \dots, X_n\}$ of discrete 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 recursive product decomposition

$$P(x_1, \dots, x_n) = \prod_i P(x_i \mid \mathbf{pa}_i) \quad (16)$$

where $\mathbf{pa}_i$ are the direct predecessors (called *parents*) of X_i in G . For example, the DAG in Figure 3 induces the decomposition

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1) P(x_2|x_1) P(x_3|x_1) P(x_4|x_2, x_3) P(x_5|x_4) \quad (17)$$

Figure 3:
A typical directed acyclic graph (DAG) representing the decomposition of Eq. (17).

The recursive decomposition in Eq. (16) implies that, given its parent set $\mathbf{pa}_i$, each variable X_i is conditionally independent of all its other predecessors $\{X_1, X_2, \dots, X_{i-1}\} \setminus \mathbf{pa}_i$. Using Dawid's [1980] notation, we can state this set of independencies as follows:

$$X_i \perp\!\!\!\perp \{X_1, X_2, \dots, X_{i-1}\} \setminus \mathbf{pa}_i \mid \mathbf{pa}_i, \quad i = 2, \dots, n \quad (18)$$

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

words, invariance is what the investigator must already have in mind when he specifies the restrictions that form the model.

preceding state. For example, the DAG of Figure 3 implies the following Markovian independencies:

$$X_2 \perp\!\!\!\perp \{0\} \mid X_1, \quad X_3 \perp\!\!\!\perp X_2 \mid X_1, \quad X_4 \perp\!\!\!\perp X_1 \mid \{X_2, X_3\}, \quad X_5 \perp\!\!\!\perp \{X_1, X_2, X_3\} \mid X_4 \quad (19)$$

Conversely, any list L of Markovian independencies defines a DAG G and permits us to decompose any P that satisfies L into a recursive product, as in Eq. (16). Normally, however, such a decomposition would imply additional independencies that are not explicit in L . For example, the decomposition of Eq. (17) implies $X_5 \perp\!\!\!\perp X_3 \mid \{X_1, X_4\}$ (which is not part of L in Eq. (19)) regardless of the numerical values assigned to the factors of that product. A graphical criterion called d -separation [Pearl 1988] permits us to read off the DAG the sum total of *all* independencies implied by a given decomposition.

Definition 4 (*d-separation*) *Let X , Y , and Z be three disjoint subsets of nodes in a DAG G , and let p be any path between a node in X to a node in Y . (By a “path” we mean any succession of arcs, regardless of their directions.) Z is said to block p if there is a node w on p satisfying one of the following two conditions:*

1. *w has converging arrows (along p) and neither w nor any of its descendants are in Z , or*
2. *w does not have converging arrows (along p) and w is in Z .*

Z is said to d -separate X from Y , in G , denoted $(X \perp\!\!\!\perp Y \mid Z)_G$, iff Z blocks every path from a node in X to a node in Y . $\square$

In Figure 3, for example, $X = \{X_2\}$ and $Y = \{X_3\}$ are d -separated by $Z = \{X_1\}$; the path $X_2 \leftarrow X_1 \rightarrow X_3$ is blocked by $X_1 \in Z$, while the path $X_2 \rightarrow X_4 \leftarrow X_3$ is blocked because X_4 and all its descendants are outside Z . Thus $(X_2 \perp\!\!\!\perp X_3 \mid X_1)_G$ holds in G . However, X and Y are not d -separated by $Z' = \{X_1, X_5\}$ because the path $X_2 \rightarrow X_4 \leftarrow X_3$ is rendered unblocked by virtue of X_5 , a descendant of X_4 , being in Z . Consequently, $(X_2 \perp\!\!\!\perp X_3 \mid \{X_1, X_5\})_G$ does not hold in G . Conceptually, 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 converging arrows at X_4 .

The d -separation criterion has been shown to be both necessary and sufficient relative to the set of distributions that are represented by a DAG G [Verma 1990, Geiger & Pearl 1988]. In other words, there is a one-to-one correspondence between the set of independencies implied by the recursive decomposition of Eq. (16) and the set of triples (X, Z, Y) that satisfy the d -separation criterion in G . Furthermore, the d -separation criterion can be tested in time linear in the number of edges in G . Thus, a DAG can be viewed as an efficient scheme for representing the Markovian independence assumptions embodied in a causal model and for deducing and displaying all the logical consequences of such assumptions.

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

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 $\langle M, P \rangle$, our main problem will be to facilitate the syntactic derivation of expressions of the form $P(x_j | \text{set}(x_i))$ 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 as follows:

Theorem 1 *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 sets of variables X , Y , Z , and W , we have:*

Rule 1 *Insertion/deletion of observations*

$$P(y | \hat{x}, z, w) = P(y | \hat{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}}} \quad (20)$$

Rule 2 *Action/observation exchange*

$$P(y | \hat{x}, \hat{z}, w) = P(y | \hat{x}, z, w) \quad \text{if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}\underline{Z}}} \quad (21)$$

Rule 3 *Insertion/deletion of actions⁵*

$$P(y | \hat{x}, \hat{z}) = P(y | \hat{x}) \quad \text{if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}\underline{Z}}} \quad (22)$$

□

Each of the inference rules above can be proven [Pearl 1994] from the basic interpretation of the “ $\text{set}(x)$ ” 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

⁵The general rule, for arbitrary conditioning set W , reads:

Rule 3' *Insertion/deletion of actions*

$$P(y | \hat{x}, \hat{z}, w) = P(y | \hat{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}, \overline{Z(W)}}} \quad (22')$$

where $Z(W)$ is the set of Z nodes that are not ancestors of any W node in $G_{\overline{X}}$. This generalization, however, is not needed for the discussions in this paper.

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 $Z = z$. It is equivalent to Eq. (11) in [Pearl 1993c], where it was named the “back-door criterion”,⁶ and 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 $q: P(y_1, \dots, y_k | \hat{x}_1, \dots, \hat{x}_m)$ 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 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 1 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.

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

$$P(z|\hat{x}) = P(z|x) \tag{24}$$

⁶The back-door criterion states that, if there exists a set S of observed variables which are non-descendants 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

$$P(y|\hat{x}) = \sum_s P(y|x, s)P(s) \tag{23}$$

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

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:

$$P(y|\hat{z}) = \sum_x P(y|x, \hat{z})P(x|\hat{z}) \quad (25)$$

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:

$$P(x|\hat{z}) = P(x) \text{ if } (Z \perp\!\!\!\perp X)_{G_{\overline{Z}}} \quad (26)$$

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

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

and note that X d -separates Z from Y in $G_{\underline{Z}}$. This allows us to write Eq. (25) as

$$P(y|\hat{z}) = \sum_x P(y|x, z)P(x) = E_x P(y|x, z) \quad (28)$$

which is a special case of the “back-door” formula (see footnote 6, (23)) 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

$(Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}}$ appears under a variety of forms, all requiring conditional-independence judgments involving counterfactual variables. The causal calculus facilitated by Theorem 1 replaces such judgments with formal tests (d -separation) on a single graph (G) representing the causal knowledge of the investigator.

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

$$P(y|\hat{x}) = \sum_z P(y|z, \hat{x})P(z|\hat{x}) \quad (29)$$

we see that the term $P(z|\hat{x})$ was reduced in Eq. (24) 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

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

since Figure 3 shows

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

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

$$P(y|z, \hat{x}) = P(y|\hat{z}) \quad (31)$$

which was calculated in Eq. (28). Substituting Eqs. (28), (31), and (24) back into Eq. (29) yields

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \quad (32)$$

Eq. (32) was named the “mediating variables” formula in [Pearl 1993b], where it was derived by direct algebraic evaluation of q_4 (taking the expectation over u) in Eq. (15).

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. (24) and (31), from which we obtain

$$\begin{aligned} P(y, z|\hat{x}) &= P(y|\hat{z})P(z|x) \\ &= P(z|x) \sum_{x'} P(y|x', z)P(x') \end{aligned} \quad (33)$$

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

$$\begin{aligned} P(x, y|\hat{z}) &= P(y|x, \hat{z})P(x|\hat{z}) \\ &= P(y|x, z)P(x) \end{aligned} \quad (34)$$

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. (26)).

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 $b = E(Y|\hat{x})/x$ is not identifiable. However, adding an arc $Z \rightarrow X$ to the structure (that is, finding a variable Z that is correlated with X but not with U) would facilitate the computation of $E(Y|\hat{x})$ via the instrumental-variable formula [Bowden & Turkington 1984]:

$$b = \frac{E(Y|\hat{x})}{x} = \frac{E(Y|z)}{E(X|z)} = \frac{R_{yz}}{R_{xz}} \quad (35)$$

In nonparametric models, adding an instrumental variable Z to a bow-pattern does not permit the identification of $P(y|\hat{x})$; one can calculate upper and lower bounds on $P(y|\hat{x})$ [Robins 1989, Manski, 1990, Balke & Pearl 1993] and these bounds may coincide for special distributions $P(x, y, z)$, but there is no way of computing $P(y|\hat{x})$ for *every* distribution $P(x, y, z)$.

A general feature of nonparametric models is that the addition of arcs to a causal diagram can impede, but never assist, the identification of causal effects. This is because such addition reduces the set of d -separation conditions carried by the diagram and, hence, if a causal effect derivation fails in the original diagram, it is bound to fail in the augmented diagram as well. Conversely, any causal effect derivation that succeeds in the augmented diagram (by a sequence of symbolic transformations, as in Corollary 1) would succeed in the original diagram.

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

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).

Figure 6:

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

joint action is often easier than computing the effects of its constituent singleton actions.⁷ 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 $G_{\overline{XZ_2}}$, giving

$$\begin{aligned} P(y|\hat{x}, \hat{z}_2) &= \sum_{z_1} P(y|z_1, \hat{x}, \hat{z}_2) P(z_1|\hat{x}, \hat{z}_2) \\ &= \sum_{z_1} P(y|z_1, x, z_2) P(z_1|x) \end{aligned} \quad (36)$$

However, Rule 2 cannot be used to convert $P(z_1|\hat{x}, z_2)$ into $P(z_1|x, z_2)$ because, when conditioned on Z_2 , X and Z_1 are d -connected in $G_{\underline{X}}$ (through the dashed lines). We conjecture, however, that whenever $P(y|\hat{x}_i)$ is computable for every singleton variable X_i , then $P(y|\hat{x}_1, \hat{x}_2, \dots, \hat{x}_l)$ is computable as well, for any subset of variables $\{X_1, \dots, X_l\}$.

⁷This was brought to my attention by James Robins, who has worked out many of these computations in the context of sequential treatment management. Eq. (36) for example, is a special case of Robin's G -computation algorithm [Robins 1989, Robins et al. 1992].

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. (14) 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.

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.

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.
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. (23)):

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

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. (33). 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

$$P(y|\hat{x}) = \sum_{z_1, z_2} P(y|z_1, z_2, x) P(z_1|x) \sum_{x'} P(z_2|z_1, x') P(x') \quad (38)$$

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

$$P(y|\hat{x}) = \sum_{z_1} \sum_{z_2} \sum_{x'} P(y|z_1, z_2, x')P(x')P(z_1|z_2, x)P(z_2) \quad (39)$$

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

(a) Y (b) Y (c) Y (d) Y

the literature on Z 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 Z 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. (37) (see Eqs. (32), (38), and (39)).

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 1). 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 .⁸ 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. (23)). However, if Z can be controlled by

⁸The 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 7 (b); it includes any set of variables that would permit (if randomized) the identification of $P(y|\hat{x})$.

randomized trial, then we can measure $P(x, y|\hat{z})$, from which we can compute $P(y|\hat{x})$ using

$$P(y|\hat{x}) = P(y|x, \hat{z}) = P(y, x|\hat{z})/P(x|\hat{z}) \quad (40)$$

The validity of Eq. (40) can be established by first applying Rule 3 to add $\hat{z}$,

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

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

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

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

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

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. (40) 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.

Note that, according to Eq. (40), 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. (40) 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. 1993]

$$E(Y|\hat{x}_1) - E(Y|\hat{x}_2) = \frac{E(Y|z_1) - E(Y|z_2)}{E(Y|x_1) - E(Y|x_2)}$$

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. (38), 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. (31), 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, in turn, 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 conflicting interpretations. 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 independent 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) 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 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 . Moreover, it is easy to show (using Rule 2) that holding the parents of Y constant has the same effect on $P(y)$ as observing those parents attain that constant. Thus, we obtain an equivalent definition for direct effect:

Corollary 2 *The direct effect of X on Y is given by $P(y|\hat{x}, \mathbf{pa}(Y)\setminus X) = P(y|x, \mathbf{pa}(Y)\setminus X)$ where $\mathbf{pa}(Y)$ are the variables appearing in the equation for Y .*

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

$$\frac{E(Y|x, \mathbf{pa}(Y)\setminus X) - E(Y|x', \mathbf{pa}(Y)\setminus X)}{x - x'}$$

reduces to the ordinary path coefficient between X and Y , regardless of the value taken by $\mathbf{pa}(Y)\setminus X$. In general, if X does not appear in the equation for Y , then $P(y|x, \mathbf{pa}(Y)\setminus X)$ defines a constant distribution on Y , independent of x , which matches our understanding of “having no direct effect”.

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 way of selectively disabling a direct 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

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

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 are necessary [Dean & Kanawaza 1989, Pearl 1993a, Balke & Pearl 1994].

4 Extensions and Limitations

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 and equilibrium equations 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 equilibrium constraint $q_d = q_s$ (connoting: quantity demanded = 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 1, 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 nonrecursive models. The subtheory interpretation of control queries (Definition 3) carries over to nonrecursive 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. (28))

$$P(y|\hat{x}) = \sum_z P(y|x, z)P(z) \quad (41)$$

will be converted to

$$E(Y|\hat{x}) = \int_y \int_z y f(y|x, z) f(z) dy dz = R_{yx \cdot z} x \quad (42)$$

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. (32) can be converted to a product

$$E(Y|\hat{x}) = R_{xz} \beta_{zy \cdot x} x \quad (43)$$

where $\beta_{zy \cdot x}$ is the standardized regression coefficient [Pearl 1993b]. More sophisticated estimation techniques, tailored specifically for causal inference can be found in [Robins 1992].

Acknowledgment

This investigation benefitted from discussions with Joshua Angrist, Peter Bentler, David Cox, 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.

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. 1993] Angrist, J.D., Imbens, G.W., and Rubin, D.B., "Identification of causal effects using instrumental variables," Department of Economics, Harvard University, Cambridge, MA, Technical Report No. 136, June 1993.
- [Balke & Pearl 1993] Balke, A., and Pearl, J., "Nonparametric bounds on causal effects from partial compliance data," UCLA Cognitive Systems Laboratory, Technical Report R-199, 1993. Submitted to the *Journal of the American Statistical Association*.
- [Balke & Pearl 1994] Balke, A., and Pearl, J., "Counterfactual Probabilities: Computational Methods, Bounds, and Applications," UCLA Computer Science Department, Technical Report (R-213-B), May 1994. To appear in the *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-94)*, Seattle, WA, July 29-31, 1994

- [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, MA, 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 Sosialokonomiske Institutt, 1948). Cited in M.S. Morgan, *The History of Econometric Ideas*, Cambridge University Press, Cambridge, MA, 1990.
- [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, 1988, pp. 136-147. Also in L. Kanal et al. (Eds.), *Uncertainty in Artificial Intelligence*, **4**, Elsevier Science Publishers, North-Holland, Amsterdam, pp. 3-14, 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.
- [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**, Yale University Press, New Haven, 1953, Chapter 1.
- [Manski, 1990] Manski, C.F., “Nonparametric bounds on treatment effects,” *American Economic Review, Papers and Proceedings*, **80**, 319-323, 1990.

- [McDonald 1993] McDonald, R., “Some problems in the application of path analysis” (in preparation).
- [Pearl 1988] Pearl, J., *Probabilistic Reasoning in Intelligence Systems*, Morgan Kaufmann, San Mateo, CA, 1988. Revised 2nd printing, 1992.
- [Pearl 1993a] Pearl, J., “From Conditional Oughts to Qualitative Decision Theory,” in D. Heckerman and A. Mamdani (Eds.), *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, 12-20, 1993.
- [Pearl 1993b] Pearl, J., “Mediating instrumental variables,” Computer Science Department, UCLA, Los Angeles, CA, Technical Report R-210, 1993.
- [Pearl 1993c] Pearl, J., “Aspects of graphical models connected with causality,” *Proceedings of the 49th Session of the International Statistical Institute: Invited papers*, Florence, Italy, August 1993, Tome LV, Book 1, pp. 391–401. Short version in *Statistical Science*, 8(3), 266-269, 1993.
- [Pearl 1993d] Pearl, J., “A note on testing exogeneity of instrumental variables,” Computer Science Department, UCLA, Los Angeles, CA, Technical Report R-211-S, 1994.
- [Pearl 1994] Pearl, J., “Causal diagrams for experimental research,” UCLA Computer Science Department, Technical Report (R-218-L), May 1994. To appear in *Biometrika*.
- [Pearl & Verma 1991] Pearl, J., and Verma, T., “A theory of inferred causation,” in J.A. Allen, R. Fikes, and E. Sandewall (Eds.), *Principles of Knowledge Representation and Reasoning: Proceeding of the Second International Conference*, Morgan Kaufmann, San Mateo, CA, 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, 1990, pp. 67-87.
- [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 Modelling*, Vol. 7, 1393-1512, 1986.
- [Robins 1989] Robins, J.M., “The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies,” in L. Sechrest, H. Freeman, and A. Mulley (Eds.), *Health Service Research Methodology: A Focus on AIDS*, NCHSR, U.S. Public Health Service, 1989, pp. 113-159.
- [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*, Vol. 3, Number 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,” *Biometrika*, 70, 41-55, 1983.
- [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 Schienens, 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, 1990, 69-76.
- [Wermuth 1993] Wermuth, N., “On block-recursive regression equations” (with discussion), *Brazilian Journal of Probability and Statistics*, 6, 1-56, 1992.